UNIVERSITY OF SOUTHAMPTON

FACULTY OF LAW, ARTS AND SOCIAL SCIENCES
SCHOOL OF SOCIAL SCIENCES

Correcting for Measurement Error when Estimating Pay Distributions from Household Survey Data

by
Gabriele Beissel-Durrant

Thesis for the Degree of Doctor of Philosophy

Division of Social Statistics

December 2003

UNIVERSITY OF SOUTHAMPTON

ABSTRACT

FACULTY OF LAW, ARTS AND SOCIAL SCIENCES
SCHOOL OF SOCIAL SCIENCES
DIVISION OF SOCIAL STATISTICS

Doctor of Philosophy

CORRECTING FOR MEASUREMENT ERROR WHEN ESTIMATING PAY DISTRIBUTIONS FROM HOUSEHOLD SURVEY DATA

by Gabriele Beissel-Durrant

The aim of this thesis is to develop and to evaluate different methods for estimating distributions in the presence of measurement error and missing data with a primary focus on a specific application concerning pay. Different methods for correcting for measurement error in a fully observed variable are considered by taking into account information on the accurately measured variable observed on a non-random subsample. To compensate for nonresponse in the correct variable and to effectively correct for measurement error in the erroneously observed variable several imputation methods are proposed treating the problem of measurement error as a missing data problem. Based on the assumption that the data are missing at random (MAR) hot deck imputation within classes as a form of predictive mean matching imputation is evaluated theoretically and empirically. This method provides approximately unbiased estimates of the parameter of interest, the proportion below a given threshold. The problem of estimating the variance of the estimator under this imputation method is investigated. A variance estimator is proposed which allows for uncertainty due to imputation. It is shown that this estimator is approximately unbiased under certain conditions.

Since some evidence is found that the results under hot deck imputation within classes may depend on the choice of imputation classes other forms of predictive mean matching imputation are evaluated theoretically and empirically under the assumption of MAR. The imputation methods are also compared to propensity score weighting. The use of repeated imputation shows gains in efficiency in comparison to single value imputation. It is found that nearest neighbour imputation using repeated imputation shows advantages in terms of bias robustness and efficiency of the point estimator. It is therefore recommended for practical use.

Several estimation methods under nonignorable nonresponse are considered making an alternative assumption of common measurement error (CME). An imputation method using data augmentation based on the assumption of CME rather than MAR is derived, which shows desirable properties of the point estimator of interest. The use of hot deck imputation in the data augmentation procedure is proposed. Data augmentation using nearest neighbour imputation under the assumption of CME is found to have desirable properties for the pay application.

Acknowledgements

This research was financially supported by the Economic and Social Research Council through studentship R42200034286.

A very big thank you goes to my supervisor, Professor Chris Skinner, for his guidance, support and encouragement throughout my PhD studies. I would like to thank the department of Social Statistics for financial support throughout my PhD and in particular for enabling me to go to two international conferences. I would like to thank all members of staff for helpful and encouraging comments, in particular Professor Ray Chambers. The friendly, research orientated environment has been very beneficial.

I would also like to thank the Office for National Statistics for the provision of datasets of the UK Labour Force Survey and interesting research topics.

Parts of the research in this thesis have been presented at conferences, such as the DataClean conference in Jyvaskyla (Finland) in May 2002 and the 54th Session of the ISI in Berlin in August 2003.

A big thank you goes to Nikos Tzavidis and Fatima Salgueiro, who I shared an office with almost throughout my entire time, as well as all other PhD students for their support and encouragement, in particular I would like to mention Marcel, Ibrahim, Leslie, Johanna, Gail, Faiza, Priscilla, Dulce and Felix. I would also like to say thank you to all the visitors of the department, in particular Pascal Riviére, Rob Clark, Pedro Nascimento Silva and Jukka.

A final thank you is due to my family, in particular my husband Paul for his patience and encouragement, and my parents who have always supported my studies.

Contents

List of Tables

List of Figures

Chapter 0

Introduction

This chapter provides a brief outline of the thesis and introduces the research problem. We shall postpone giving references to later chapters. Distributions of hourly pay are important for a wide range of social and economic policy issues. Of considerable interest is how policy interventions like minimum wage laws impact on such pay distributions. Examples are the analysis of low pay and its relation to poverty or comparisons between high and low-income earners and the analysis of income inequality. The investigation of low pay is of particular interest to a wide range of analysts in economics and the social sciences in Great Britain due to the introduction of the National Minimum Wage (NMW) there in April 1999. This legislation received a lot of interest from a wide range of organisations, unionists, employers, economists, politicians and media. It is thought that this policy has effects on for example earnings, pay distributions, poverty and income inequality. To analyse the effects of this new law it is crucial to have reliable data about the hourly pay of employees in the UK, especially for the bottom end of the pay distribution. However, it is difficult to obtain reliable data on both earnings and hours since such variables are often prone to nonsampling errors, such as nonresponse and measurement error. The primary aim of this dissertation is to develop methods for improving estimates of pay distributions based upon large household survey data. While this thesis is driven by a specific applied problem, methods are developed and evaluated in a general framework so that in principle they would be applicable to other applications.

We use data from the UK Labour Force Survey, a large survey of households, which includes information on hours worked and earnings of employees. This survey is conducted quarterly by the UK Office for National Statistics (ONS). Given this information about hours and earnings it is possible to derive a variable measuring the hourly pay of employees. However, this derived variable appears to be subject to a considerable amount of measurement error, which may lead to an overestimation of the lower end of the pay distribution. An alternative variable on hourly earnings is obtained by asking employees directly about their hourly pay. This direct variable appears to give very accurate information but is subject to a high amount of missing data, since it only applies to employees that are paid on an hourly wage basis. Therefore, many individuals are not able to report their hourly pay. In a sense, this direct variable can be regarded as internal validation data obtained on a non-random subsample of the whole survey. The aim is to use both variables, the direct and the derived variable, for estimating the distribution of hourly earnings.

Systematic but also random measurement error can lead to serious bias, when estimating distributional quantities, such as quantiles and proportions, particularly in the tails of the distribution. A major aim is to correct for this bias in the derived variable. Due to the missing data in the direct variable the aim is to impute the missing values taking into account information on the erroneous variable and other covariates, such that the imputation method effectively corrects for the measurement error in the pay variable and compensates for potential distorting effects. The focus therefore is on correcting for measurement error and adjusting for nonresponse in pay variables.

In chapter one the relevant literature on measurement error with particular focus on response error in income and earnings variables is reviewed. In addition, definitions and terminology important in the missing data context are introduced. Several imputation methods are reviewed with particular emphasis on nearest neighbour, hot deck, regression, predictive mean matching and multiple imputation.

The second chapter describes the data available on earnings in the UK with emphasis on the Labour Force Survey (LFS). Of interest is the sampling design of the LFS and the type of earnings variables available in this dataset. The initial imputation methodology carried out by the ONS is reviewed, which under the missing at random assumption (MAR) uses a form of predictive mean matching imputation. It is a random hot deck procedure within imputation classes based on a

regression model. The imputation is applied multiple times. Based on this imputation method the aim is to estimate the distribution of pay. The point estimator of interest, which is the proportion of employees earning below or around the NMW, is introduced and estimates of pay distributions based on the imputed direct variable are presented.

Chapter three evaluates hot deck imputation within classes. Theoretical properties of the point estimator are investigated, also taking into account weighting in the LFS. A simulation study is carried out to evaluate the performance of the point estimator. Limitations and advantages of this imputation method, in particular robustness against model misspecification and dependency of resulting estimates on the choice of classes, are investigated.

In chapter four the variance of the point estimator of interest under hot deck imputation within classes is investigated which allows for uncertainty due to imputation. A formula for variance estimation taking into account imputation, response and sampling variability as well as allowing for fixed survey weights is derived using a design-based approach. In addition, variance estimation using Rubin's multiple imputation formula is considered. Because the imputation method does not constitute 'proper' multiple imputation, this approach is shown to underestimate the variance. An adjusted multiple imputation formula is derived and shown to provide approximately unbiased variance estimation. A simulation study is carried out to compare the performances of several variance estimators. For comparison the approximate Bayesian bootstrap is implemented and variance estimation is carried out using Rubin's multiple imputation variance formula.

In chapter five a wider class of predictive mean matching imputation methods is investigated, including nearest neighbour imputation using repeated imputation, stochastic and deterministic imputation as well as imputation using a penalty function. The performance of the point estimator of interest under these imputation methods is investigated theoretically and empirically with particular emphasis on bias robustness, efficiency and robustness against model misspecification. The imputation methods are also compared to propensity score weighting as an alternative way of compensating for nonresponse bias. It is found that nearest neighbour imputation using repeated imputation performs slightly better in terms of robustness and efficiency than hot deck imputation within classes and propensity score weighting. It is therefore recommended for practical use.

Chapter six focuses on the measurement error in the derived variable of hourly earnings. The measurement error is modelled to evaluate the validity of classical measurement error assumptions. It is found that many of these assumptions are violated such that standard measurement error approaches cannot be used to correct for measurement error in the derived variable. An alternative assumption of the nonresponse mechanism is proposed referred to as the common measurement error assumption (CME), which allows the nonresponse to be dependent on the variable subject to missing data. The main focus of this chapter is the development of imputation and estimation methods valid under such nonignorable nonresponse. Several methods using either deconvolution, misclassification and a weighted bootstrap approach in a Bayesian framework are investigated under the CME assumption. Data augmentation using rejection sampling valid under the common measurement error assumption, and therefore under nonignorable nonresponse, is developed, showing desirable properties of the point estimator of interest. The use of hot deck imputation as an alternative to parametric regression imputation in the data augmentation procedure is proposed. A simulation study is carried out to evaluate the method in terms of bias and efficiency. This CME-based method is compared to the previously considered MAR-based imputation methods. An application of different imputation methods to the LFS is considered.

Chapter seven summarises the main conclusions and highlights the main contributions of the thesis. Possible areas of further research are discussed.

Chapter 1

Review of Literature on Measurement Error, Missing Data and Imputation

As background to the problem of estimating earnings distributions based on variables that are prone to measurement error and missing data, as described in the introduction, the literature on methods correcting for measurement error and compensating for nonresponse bias is reviewed. Measurement error occurs if a recorded value on a study variable for a sampled element differs from the true value, assuming a true value exists. Nonresponse occurs if a whole unit does not respond or different items in the survey are subject to missing data (Särndal, Swensson and Wretman, 1992; Lessler and Kalsbeek, 1992).

To facilitate our discussion we introduce the following notation. Let Y denote the variable of interest measured without error, X the variable measured with error, I a binary indicator of whether Y is observed and W a vector of auxiliary variables. Let U be a finite population of N units and s a sample of sample size n. The values y_i, x_i, I_i and w_i, where $i = 1,...,n$, are the sample values of the variables Y, X and I and of a $(1\times V)$-vector of variables W of the form $W = (W_1,...,W_V)$. The vectors of length n, containing the sample values are denoted $\tilde{Y}$, $\tilde{X}$ and $\tilde{I}$ such that $\tilde{Y} = (y_1,..., y_n)'$, $\tilde{X} = (x_1,...,x_n)'$ and $\tilde{I} = (I_1,...,I_n)'$, and $\tilde{W}$ denotes a matrix with

values of the covariates. For the variable Y only the first n_r elements are observed in sample s and the following $n-n_r$ elements are missing. We have $\tilde{Y}=(\tilde{Y}'_{obs},\tilde{Y}'_{mis})'$, where $\tilde{Y}_{obs}=(y_1,...,y_{n_r})'$ is the observed part of $\tilde{Y}$ and $\tilde{Y}_{mis}=(y_{n_r+1},...,y_{n-n_r})'$ the missing part. The variable I indicates if Y is observed or not such that

$$I_i=\begin{cases}1 & \text{if } y_i \text{ observed}\\ 0 & \text{if } y_i \text{ missing.}\end{cases} \tag{1.1}$$

Thus, we suppose without loss of generality that $I_1=...=I_{n_r}=1$ and $I_{n_r+1}=...=I_n=0$. We assume that the variable X, measuring Y with error, is fully observed for all units i in sample s. Further it is assumed that the covariates W are fully observed. We shall also consider a more general case where H denote a $(1\times K)$-vector of variables of interest and $\tilde{H}$ the complete data matrix with element h_{ik} in the ith row and kth column, where $i=1,...,n$ and $k=1,...,K$. In the presence of missing data $\tilde{H}_{obs}$ refers to the observed part of the matrix $\tilde{H}$ and $\tilde{H}_{mis}$ to the missing part. In this general case $\tilde{I}$ denotes a matrix with elements

$$I_{ik}=\begin{cases}1 & \text{if } h_{ik} \text{ observed}\\ 0 & \text{if } h_{ik} \text{ missing.}\end{cases} \tag{1.2}$$

We shall adopt a statistical modelling framework when the observations $\tilde{Y},\tilde{X},\tilde{I},\tilde{W}$ and $\tilde{H}$ are realised values of corresponding random vector and matrices. For simplicity, in much of the discussion of this chapter we shall assume that the (y_i,x_i,I_i,w_i) are independently and identically distributed outcomes of a random vector denoted (Y,X,I,W). Similarly we assume the rows of $\tilde{H}$ are independent and identical realisations of a random vector H. We shall use f to denote a generic probability density, for example $f(Y\mid X)$ denotes the conditional distribution of Y given X.

In section 1.1 measurement errors in surveys and their effects on statistical analysis are described. Some adjustment methods compensating for the effects of measurement error are presented. Section 1.2 focuses on measurement errors in earnings and income variables. In section 1.3 nonresponse error and different nonresponse mechanisms are discussed. A number of imputation methods as a way of handling nonresponse are presented in section 1.4, with particular emphasis on nearest neighbour, hot deck, regression and multiple imputation.

1.1 Measurement Errors in Surveys

Särndal, Swensson and Wretman (1992) describe *measurement errors* as those errors in individual data that occur during the data collection stage. They define measurement error as the difference between the recorded value on a study variable for a sampled element and the true value, which is assumed to exist. In a survey context the expression *response error* is often used instead of measurement error. When estimating distribution functions, as opposed to simple estimation such as estimation of means and totals, it is important to investigate the cause and the characteristics of such measurement errors since parameters such as quantiles can be strongly affected by measurement error even under common error assumptions. The impact of such error needs to be investigated as well as possible adjustment methods to compensate for the effects of measurement error. Often the impact and structure of the measurement error can only be analysed if true values are available for comparison, which might be difficult to obtain. In the following the literature on measurement error in surveys is reviewed, focussing on common measurement error models.

1.1.1 Measurement Error Models

In the theory of measurement error two main approaches can be distinguished. Biemer and Stokes (1991) refer to the *sampling approach* if the measurement error is viewed as a random variable sampled from a hypothetical error distribution. The observed variable is the outcome of a two-stage random sampling procedure. It is determined by sampling from a finite population and adding on a random error sampled from an infinite population of errors. The errors can occur because of misunderstandings, typing or coding errors etc. The other approach, referred to as the *psychometric approach*, aims to determine the relationship between multiple responses from a unit belonging to a sample, to evaluate the correctness of individual responses and to explain the relationship between responses and errors.

To deal with measurement error we first assume the existence of a true value of the variable of interest Y, for each individual in the population $i \in U$. The notion of a true value, however, can be controversial, which is discussed in greater detail below. This true value is denoted y_i for individual i, whereas the actual observed value in a sample is denoted x_i. We write

$$x_i = y_i + \delta_i \quad \forall\, i \in s, \tag{1.3}$$

where the difference between the observed value and the true value δ_i is called the *measurement error* for element i. The observed variable x_i is sometimes called the *manifest* or *indicator* variable, and the either unobserved or only partially observed, true variable Y the *latent* variable. Models where y_i is fixed are called *functional* models, whereas if y_i is random it is referred to as a *structural* model (Fuller, 1987; Särndal, Swensson and Wretman, 1992, Lessler and Kalsbeek, 1992; Carroll, Ruppert and Stefanski, 1995). When trying to estimate parameters that are based on variables subject to measurement error it is necessary to make certain assumptions about the measurement error and to formulate a measurement error model.

In the following a simple measurement error model making assumptions about the structure of the errors is presented. It is assumed that the measurements for each element differ for repeated measurements. The underlying assumption of the following model is that the measurements, and therefore the measurement errors, are regarded as random variables, such that standard statistical tools, for example for evaluating the precision of an estimation, can be used. We describe a model for measurements of elements from a random sample drawn from a finite population U of size N. We treat the true values as fixed and so this is a functional model. Biemer and Stokes (1991) denote this approach the 'sampling approach' (see also Särndal, Swensson and Wretman, 1992; Lessler and Kalsbeek, 1992).

Following the approach of Särndal, Swensson and Wretman (1992) drawing a sample s according to a specific sampling design D the random variables x_i, for all $i \in s$, are assumed to have a joint probability distribution, conditional on s. This is called a *measurement model*, denoted M_s or simply M. It should be noted that in many applications the dependence on s is omitted. The data generation contains two random stages with

1. Selection of a sample s according to D
2. Measurement model M, which generates an observed value x_i for all $i \in s$.

The measurement error model is formulated as in (1.3). For this model y_i is a random variable, whose distribution depends on the sample design. The error term δ_i is regarded as a single random draw from an infinite population of errors for each individual unit i. The *classical measurement error model* assumes that (Biemer and Stokes, 1991):

1. The true values y_i exist and are well-defined.
2. The measurement error is additive.

Let $E_M(.|i)$ and $\text{var}_M(.|i)$ denote the expectation and variance according to the measurement error model based on the i-th unit. It is assumed

3. $E_M(\delta_i|i)=0$ (1.4)
4. $\text{var}_M(\delta_i|i)=\sigma^2_{\delta i}$ (1.5)
5. $\text{cov}_M(\delta_i,\delta_j)=0$, for all $i \neq j$. (1.6)

It follows that $\text{cov}_{DM}(y_i,\delta_j)=0$ for all i and j, (therefore also $\text{cov}_{DM}(y_i,\delta_i)=0$, see also Saris and Andrews (1991)), which means that there is no correlation between the errors and the latent variables. According to assumption 3 we have:

$$\text{cov}_{DM}(y_i,\delta_j)=E_{DM}(y_i\delta_j)-E_{DM}(y_i)E_{DM}(\delta_j)=E_D(y_iE_M(\delta_j|j))-E_D(y_i)E_M(\delta_j|j)=0.$$

In addition, it is sometimes assumed that the error terms are i.i.d. (i.e $\text{var}_M(\delta_i|i)=\sigma^2_{\delta i}=\sigma^2_\delta$, for all i) and in particular follow a normal distribution $\delta_i \sim N(0,\sigma^2_\delta)$. To obtain variations on the classical measurement error model it is possible to relax some of the assumptions. It is also possible to think of a multiplicative model, e.g. by viewing $\delta_i = y_i(e_i-1)$ such that $x_i = y_ie_i$.

In the above model we assume the existence of a true value. However, defining and determining a true value depends very much on the variable of interest. It requires that the characteristic to be measured has a clear operational definition and that precise measurement methods can be implemented. In general, it is assumed that a true value for variables such as age, height, earnings etc. exists. For subjective phenomena such as opinions, attitudes, beliefs or concepts such as satisfaction or emotion it is difficult to define a true value (Biemer and Stokes, 1991; Lessler and Kalsbeek, 1992; Särndal, Swensson and Wretman, 1992). To handle this problem it might be assumed that there exists a response distribution for repeated measurements x_{it}, $t=1,...,2$, and that y_i is defined as the mean of the response distribution over an infinite number of trials (Fuller, 1995), such that

$$y_i = E_M(x_{it}|i). \quad (1.7)$$

The error term is defined as the difference between the observed value and this mean, such that

$$\delta_{it} = x_{it} - y_i. \quad (1.8)$$

We still have $E_M(\delta_{it} \mid i) = 0$ since y_i is defined as the mean of the response distribution. In this, as Biemer and Stokes (1991) call it, psychometric approach, the same assumptions as before are made, apart from the different interpretation of y_i as a mean of the response distribution. This model is referred to as the *'classical true score model'*. Saris and Andrews (1991) point out that the term 'true score' only refers to the observed score minus the measurement error. The actual true value of a unit on a latent variable is unknown.

An extension to the above simple measurement error model are models that incorporate interviewer effects, which in particular may introduce correlated measurements. Similar correlations can be introduced by other factors such as members belonging to the same household etc. (Särndal, Swensson and Wretman, 1992). Biemer and Stokes (1991) describe a model where assumption 5 given in (1.6) is relaxed and the measured values of two different sample units may be correlated.

1.1.2 Effects of Measurement Errors on Statistical Analysis

Measurement errors in surveys are often non-negligible and can have a considerable impact on estimates of population parameters. As already mentioned, measurement errors that follow the common assumptions usually do not affect the bias of estimates for means and totals but affect the variance (Biemer and Trewin, 1997). For other statistics such as quantiles of distributions or regression coefficients the situation is more complex and even under classical assumptions substantial bias can be introduced.

Since in this thesis the focus is on estimating the proportion of low paid employees, distribution functions of earnings and regression coefficients for modelling hourly earnings, the effects of measurement errors on proportions, quantiles and regression coefficients are reviewed. For proportions the measurement error model for binary data is briefly discussed (Biemer and Stokes, 1991; Biemer and Trewin 1997). The assumptions for continuous data described earlier are not necessarily appropriate for binary data, e.g. the assumption that $\mathrm{cov}_{DM}(y_i, \delta_i) = 0$ does not hold in general. Let the variables Y and X denote binary variables, taking values 0 or 1. We define the *misclassification probabilities* as

$$m_{01i} = P(x_i = 0 \mid y_i = 1) = P(\delta_i = -1 \mid y_i = 1) \tag{1.9}$$

$$m_{10i} = P(x_i = 1 \mid y_i = 0) = P(\delta_i = 1 \mid y_i = 0), \tag{1.10}$$

where m_{01i} is the probability of a false negative and m_{10i} is the probability of a false positive. Biemer and Stokes (1991) replace assumption 3 in (1.4) by

3'. $$E_M(\delta_i \mid i) = -y_i m_{01_i} + (1 - y_i) m_{10_i} \text{ for all } i \in s, \tag{1.11}$$

where y_i is regarded as fixed in the population. (Note that $E_M(x_i \mid i) = E_M(y_i + \delta_i \mid i)$ $= y_i(1 - m_{01_i}) + (1 - y_i)m_{10_i}$). We have $E_M(\delta_i \mid i) = 0$ if $m_{01} = m_{10} = 0$, i.e. if there is no misclassification error. When analysing the effect of measurement error we first assume that the probabilities of a false negative or a false positive do not depend on the sample unit or on the interviewer, i.e. $m_{01i} = m_{01}$ and $m_{10i} = m_{10}$ for all i. The sample estimator subject to measurement error is defined as

$$\hat{P}^* = \frac{1}{n}\sum_{i \in s} x_i \tag{1.12}$$

Biemer and Trewin (1997) show that the bias in the sample proportion is

$$\text{Bias}(\hat{P}^*) = E_{DM}(\hat{P}^*) - P = [P(1 - m_{01}) + (1 - P)m_{10}] - P = -m_{01}P + (1 - P)m_{10}. \tag{1.13}$$

Thus, $\text{Bias}(\hat{P}^*) = 0$ if either $m_{01} = m_{10} = 0$, i.e. if no measurement error exists, or if

$$\begin{aligned} & Pm_{01} = (1 - P)m_{10} \\ \Leftrightarrow \quad & P(y_i = 1)P(x_i = 0 \mid y_i = 1) = P(y_i = 0)P(x_i = 1 \mid y_i = 0) \\ \Leftrightarrow \quad & P(x_i = 0, y_i = 1) = P(x_i = 1, y_i = 0), \end{aligned} \tag{1.14}$$

i.e. if the number of false negative misclassifications in the population is exactly the same as the number of false positive misclassifications, independent of the size of P. We can see that in the presence of measurement error it is very unlikely that $\hat{P}^*$ is unbiased for P. However, the bias might be negligible if m_{01} and m_{10} are small. In the case of a small proportion P the effect on the relative bias can be large even if m_{10} is small (see example in Biemer and Trewin, 1997, p. 615).

We now focus on potential effects of measurement error on distribution functions and quantiles. The effects of measurement error on the estimation of quantiles have been considered by Fuller (1995) and Biemer and Trewin (1997) assuming the population to be sampled is normally

distributed. Let Y denote the random variable of interest and F_Y its cumulative distribution function, such that $F_Y(y) = P(Y \leq y) = a$. The a-th quantile of a distribution, y_a, is defined as

$$Q_Y(a) = \inf_y (F_Y(y) > a) = y_a . \tag{1.15}$$

The quantile function is the inverse of the distribution function, i.e. $Q_Y(a) = F_Y^{-1}(a)$. In the following we assume that the true random variable Y and the error terms follow a normal distribution. Thus, the observed values also follow a normal distribution. We are interested in the a-th quantiles according to the distribution of the observed and the true values, i.e.

$$Q_X(a) = \inf_x (F_X(x) > a) = x_a \text{ and } Q_Y(a) = \inf_y (F_Y(y) > a) = y_a , \tag{1.16}$$

such that

$$F_X(x_a) = P(X \leq x_a) = a = P(Y \leq y_a) = F_Y(y_a) . \tag{1.17}$$

Let us assume that X and Y have mean $\bar{X}$ and $\bar{Y}$ respectively. We have

$$\frac{x_a - \bar{X}}{\sigma_X} = \frac{y_a - \bar{Y}}{\sigma_Y} . \tag{1.18}$$

Let us define the ratio $R = \sigma_Y^2 / \sigma_X^2$, where $\sigma_X^2 = \sigma_Y^2 + \sigma_\delta^2$. It follows for the relationship of the a-th quantiles for the distribution of the true and the observed values that

$$Q_X(a) = x_a = \frac{\sigma_X}{\sigma_Y}(y_a - \bar{Y}) + \bar{X} = R^{-1/2}(Q_Y(a) - \bar{Y}) + \bar{X}. \tag{1.19}$$

For the centered distribution of observations, $X - \bar{X}$, we obtain

$$x_a - \bar{X} = \frac{\sigma_X}{\sigma_Y}(y_a - \bar{Y}) \tag{1.20}$$

and therefore

$$Q_{X-\bar{X}}(a) = R^{-1/2} Q_{Y-\bar{Y}}(a) . \tag{1.21}$$

We see that the a-th quantiles for the distributions of the true and the observed values are not the same in the presence of measurement error. The quantiles of $X - \bar{X}$ are multiples of the

quantiles of $Y-\bar{Y}$. Figure 1.1 illustrates the effect of measurement error on distribution functions for the example where the true values are $y_i \sim N(0,1)$ and the error terms are $\delta_i \sim N(0,1)$ for all i. Thus, the observation x_i is a $N(0,2)$ distributed random variable. Since the variance of X is greater than the variance of Y and both variables are normally distributed with mean zero, it follows that the two cumulative distribution functions are only the same at the mean of X and Y. In this example we have $Q_{Y-0}(0.95) = 1.65 = y_a$ and $Q_{X-0}(0.95) = (1/2)^{-1/2} * 1.65 = 2.33 = x_a$. The 95th percentile of the distribution of the observed values is therefore larger than the 95th percentile from the distribution of the true values.

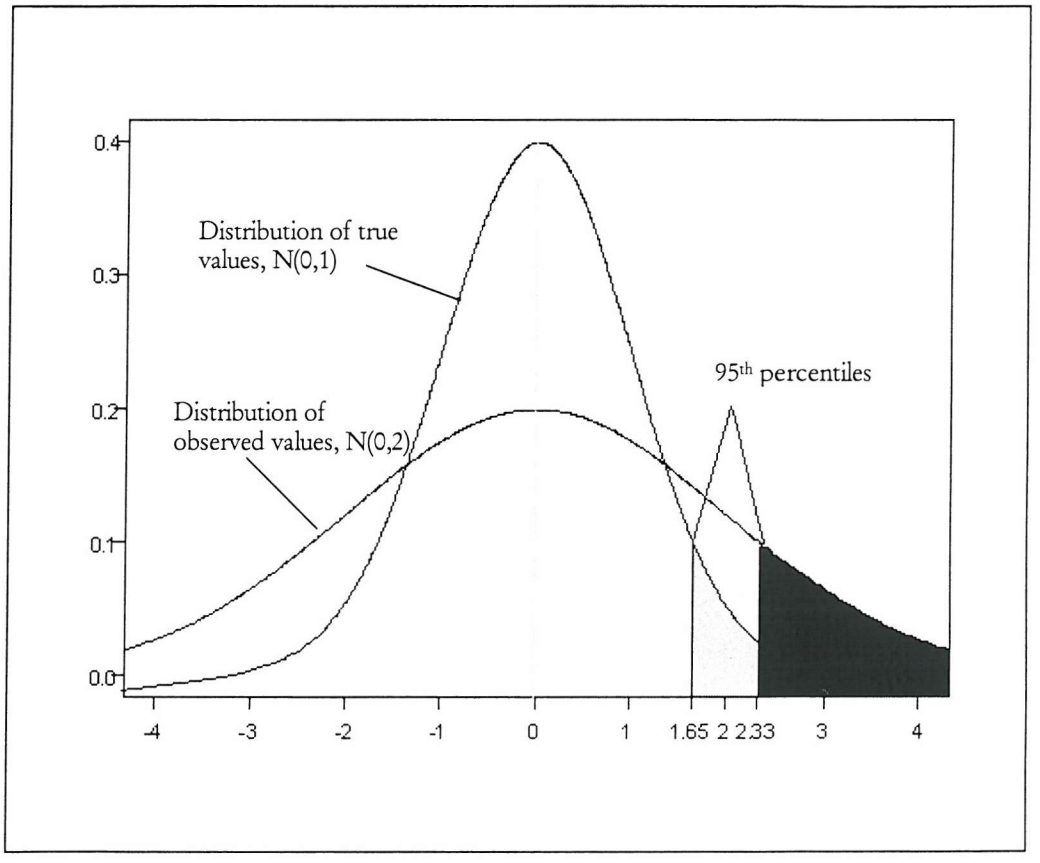

Figure 1.1: The effect of measurement error on distribution functions (Biemer and Trewin, 1997, p. 613).

The bias of the estimator $\hat{Q}_X(a)$, if $\hat{Q}_X(a)$ is an unbiased estimator of $Q_X(a)$ in the absence of measurement error, is

$$Bias(\hat{Q}_X(a)) = E_{DM}(\hat{Q}_X(a)) - Q_Y(a) = Q_X(a) - Q_Y(a)$$

$$= R^{-1/2}(Q_Y(a) - \bar{Y}) + \bar{X} - Q_Y(a) \qquad \text{(using (1.19))}$$

$$= Q_Y(a)(R^{-1/2} - 1) - R^{-1/2}\bar{Y} + \bar{X} = Q_Y(a)(R^{-1/2} - 1) - R^{-1/2}\bar{Y} + \bar{Y} + B_\delta$$

where B_δ is the measurement bias, i.e. the bias in the sample mean, $E_{DM}(\hat{\bar{X}}) = \bar{Y} + B_\delta$, where $\hat{\bar{X}}$ denotes the sample mean, such that $B_\delta = E_{DM}(\hat{\bar{X}}) - \bar{Y} = \bar{X} - \bar{Y}$ which is zero under the classical error assumptions.

$$= (R^{-1/2} - 1)(Q_Y(a) - \bar{Y}) + B_\delta \qquad (1.22)$$

Thus, in general the sample cumulative distribution function of X is a biased estimator of the cumulative distribution function of Y. However, the effect of measurement error for cumulative distribution functions depends on the part of the distribution. For $a = 0.5$ the bias of the estimated median under the assumption of the normal distribution with mean 0 is B_δ, which is zero under common error assumptions. However, if one moves away from the median the relative error increases. For $a \geq 0.5$ the bias increases as a increases and the ratio R decreases. The effect of measurement error is therefore particularly apparent when estimating tails of the distribution. A more detailed discussion of the effects of measurement error on estimation of quantiles can be found in Fuller (1995), Biemer and Trewin (1997) and Nusser et al. (1996).

Measurement error also affects estimators of regression coefficients, sometimes even if the assumptions of the classical measurement error model hold (Fuller, 1987; Biemer and Trewin, 1997). Two cases are distinguished depending on if the dependent variable of the regression model is subject to measurement error or if one or more of the independent variables are affected. Suppose we wish to estimate the simple regression model with only one covariate Y. We have

$$z_i = \beta_0 + \beta_1 y_i + \varepsilon_i, \qquad (1.23)$$

where z_i and y_i are measured without error and ε_i is identically and independently distributed with mean zero, variance σ_ε^2 and $\text{cov}_D(y_i, \varepsilon_i) = 0$. If measurement error occurs in the dependent variable the measurement error under classical error assumptions does not bias estimates for coefficients, i.e. β_0 and β_1, but increases the variance. A particular concern is the case when the

usual measurement error assumptions do not hold. Under these circumstances, estimates of population parameters might be severely biased. If measurement error occurs in the independent variable, such that X is the measurement of true Y and X follows the measurement error model in (1.3), all estimated coefficients are biased towards zero, even if the common measurement error assumptions hold. As Fuller (1987) describes it, the estimate of the coefficient is attenuated under the measurement error. The degree of attenuation under the classical assumptions can be defined by the ratio R,

$$R = \frac{\sigma_Y^2}{\sigma_X^2}. \tag{1.24}$$

In the presence of measurement error the expected value of the least squares estimator of β_1 is

$$E_m(\hat{\beta}_1) = \beta_1 \frac{\sigma_Y^2}{\sigma_X^2}, \tag{1.25}$$

where the estimate $\hat{\beta}_1$ is based on the observed values x_i. This enables us to find an unbiased estimator for β_1 in the presence of measurement error if R is known. Biemer and Trewin (1997) also discuss the case where measurement errors are correlated both within variables and between variables. In the presence of correlated error estimates of the coefficients may be substantially overestimated or underestimated depending on the signs and magnitudes of the correlations. Detailed analysis of the effect of measurement error in regression models is given in Fuller (1987). He concentrates on the case where the independent variable is not observed correctly, taking into account if only one or several explanatory variables are used in the regression. Rodgers and Herzog (1987) discuss the case of a regression model, where one variable is subject to measurement error, in particular the dependent variable. They show that violation of the common measurement error assumptions introduces bias into the standard ordinary least squares estimates (see also Duncan and Hill, 1985; Fuller, 1987; Brownstone and Valletta, 1996).

1.1.3 Methods Compensating for the Effects of Measurement Error

As we have seen, measurement error can have substantial effects on estimates of population parameters. Measurement error can lead to biased results and to an inflation of the variance of an

estimator even under classical measurement error assumptions. Of particular interest therefore are possible adjustment methods to correct for measurement error. To adjust for the bias introduced by measurement error either information about the type and magnitude of the measurement error is needed or certain assumptions about the error terms need to be made. Biemer and Trewin (1997) characterize adjustment methods by two requirements: a.) the assumed underlying model for the error structure and b.) auxiliary data for estimating the error parameters indicated in the model. For a comprehensive review of measurement error problems and adjustment methods see Carroll, Ruppert and Stefanski (1995).

Kuha and Skinner (1997) distinguish two kinds of data to correct for potential distorting effects of measurement error. The first type is *validation data*, which is assumed to include true values. Usually it is too expensive to obtain accurate information on the whole sample s, so that correct but more expensive and time involving measures can only be obtained for a smaller sample s_v. We refer to the term *internal validation data*, if the validation sample s_v is a subsample of the sample s, obtained by a known randomised double sampling scheme, i.e. the subsample s_v is a random sample of s such that s_v is representative of s. The disadvantage of such an approach is that some units in the sample may need to be interviewed twice. Another possibility is to obtain *external validation data*, for example by using company or tax records to compare survey responses with information from this external source. The company or tax records are usually assumed to give accurate information.

Several methods have been developed to compensate for measurement error in the presence of validation data obtained by a randomised double sampling scheme. Rao and Sitter (1997) propose the use of validation data to adjust estimates of means and totals for the measurement bias. The methods are based on a two-phase sampling approach, i.e. correct measurements are obtained on a random subsample of the initial sample. Luo, Stokes and Sager (1998) propose methods for estimating cumulative distribution functions in the presence of a two-phase sample. Several estimators are compared, such as the ratio, difference and regression estimator for a two-phase sample and two proposed estimators based on a weighted average of perfect and imperfect measurements. Buonaccorsi (1990) considers the use of a wider class of measurement error models, other than the simple additive one, relating true and observed values. These methods require the availability of external or internal data containing true values to validate the measurement error model. Correcting for measurement error is based on double sampling, where

true values are obtained on a random subsample. Maximum likelihood estimators are developed for certain models making assumptions about the nature of the measurement error. Another method for estimating distribution functions in the presence of measurement error is the SIMEX method (simulation - extrapolation) discussed by Carroll, Ruppert and Stefanski (1995) and Stefanski and Bay (1996). This method makes assumptions about the distribution of the error term, such as assumptions of additive and normal measurement error. An estimate of the variance of the error terms is required which can be obtained from a two-phase sample.

When true values are available for a subset of the original sample, one important approach is to view the problem of correcting for measurement error as a missing data problem, regarding the true values as missing data in the remaining sample. Note that the validation data does not necessarily need to be obtained as a random subsample of the original sample. Methods developed to compensate for nonresponse, such as imputation and weighting, can then be used to correct for measurement error. For example imputation methods can be used to fill in the missing data in the correctly measured variable. Such an approach has been proposed by Kuha (1997). He uses data augmentation to fill in missing values in the true variables to correct for measurement error in estimates of regression coefficients. Carroll, Ruppert and Stefanski (1995) also view measurement error models as special kinds of missing data problems. More on the use of imputation methods can be found in section 1.4.

The second type of data described in Kuha and Skinner (1997) is obtained by using *repeated measurements* on the variables subject to measurement error. This is particularly useful if a true value of the characteristic of interest is not well-defined or if it is not possible to measure the true value, even on a small sample of limited size. To obtain repeated measurements *reinterviews* can be carried out, for example on a subsample of the whole sample. It is also possible to ask the same or a similar question twice in an interview or questionnaire. In this approach it is often assumed that the measurements are independent of each other given the true values. However, it is possible to make assumptions about the relationships between the observed values and the latent variables and between the latent variables themselves. Data based on repeated measurements may therefore be analysed using latent variable models. Kuha and Skinner (1997) summarize methods for correcting for measurement error in categorical data using either validation data or repeated measurements. Fuller (1995) proposes a method of estimating distribution functions and quantiles

assuming that replicated measurements on some units in the sample are available. He gives a consistent estimator of the quantile function and a formula for its variance estimation using estimators of the error variances of the measurement error. An example is discussed based on data from a survey on food intakes where four replicates are available for some units. Other examples for estimating distribution functions in the presence of measurement error using repeated measurements are discussed in Nusser et al. (1996) and Fuller and Guenther (1997) who estimate the distribution of usual food intake, where usual food intake is an example of a latent variable that cannot be observed directly. A feasible way of obtaining the required information is to ask people about their eating habits on several days throughout a longer period of time and to use the average of these figures to get an indicator of the true but unobserved values. It should be noted that a considerable amount of literature exists describing the use of validation data. Often it is not possible to carry out reinterviews or to have access to a reliable external source of data. In these circumstances quantifying the measurement error is a difficult task and relies mainly on unproven assumptions.

1.2 Measurement Error in Income and Earnings Variables

Since income data is important for a wide range of policy issues, the quality of survey measures of income is of particular interest. Recent survey validation studies suggest that measurement error in income and earnings data is a common problem and leads to bias in survey estimates. In particular, variables of hourly earnings are often prone to measurement error (Bound et al., 1990; Moore, Stinson and Welniak, 2000). It is therefore of interest to analyse the effects of measurement error in income and earnings variables. In statistical and economical data analysis strong assumptions are sometimes made about the measurement error structure, for example regarding the independence of the measurements. These assumptions, however, might not hold in reality, and often classical measurement error assumptions are violated. In the following sections the focus is on measurement error in earnings variables. Section 1.2.1 describes the nature of measurement error in such variables. In section 1.2.2 it is shown that common measurement error assumptions are often violated in income and earnings variables.

1.2.1 Measurement Error in Income and Earnings Variables

There are several reasons for a high amount of measurement error in variables related to income. One reason for measurement error in such variables is the fact that total income refers to a variety of different income sources. Griffiths and Wall (1999) refer to three sources of income, such as income from labour (wages and salaries), from the ownership of capital (dividend and interest) and land (rent). Atkinson (1996) includes five sources, i.e. 1.) wages and salaries, 2.) income from self-employment, 3.) rent, dividends and interest, 4.) occupational pensions and life insurance and 5.) state transfers. The factor labour, i.e. wages and salaries, usually makes up the largest part of total income. Other reasons for measurement error in income or earnings variables are often related to difficulties in understanding the terminology of the survey question or not remembering the exact amount or period of pay. Another common problem when reporting earnings is rounding or truncation errors with the result that round numbers show a higher frequency (Rodgers and Herzog, 1987; Bound et al., 1990; Nordberg, Penttilae and Sandstroem, 2001). Respondents may also deliberately give wrong answers, e.g. for tax reasons, or misreport their period of pay, their earnings during that period or hours worked. Several studies have shown that employees sometimes considerably under- or overreport their income or earnings. Employees with lower-than-average earnings are likely to overreport, whereas high-wage workers often underreport their true earnings. Moore, Stinson and Welniak (2000) find that survey respondents usually underreport income and that underreporting errors tend to predominate over overreporting errors of income. The magnitude of the underreporting is highly variable across different types of income. However, underreporting seems to affect wage and salary income, in which we are mainly interested in, only very modestly. According to Moore, Stinson and Welniak (2000) several studies confirm that only a very low net bias in survey reports of wage and salary income and only little random error exists. Pischke (1995) finds that reported earnings and earnings from validation data does not differ much in either mean or variance, which implies that the mean of the measurement error is approximately zero. Duncan and Hill (1985) report that there is a tendency for workers to significantly overreport hours worked. Another possible source of error might be interviewer effects.

Much of the research about measurement error in income and earnings variables is based on the PSIDVS validation study (Panel Study of Income Dynamics Validation Study) and the CPS (Current Population Survey) (Bound et al., 1990 and 1994; Bound and Krueger, 1991; Pischke,

1995). The PSIDVS validation study for example allows for a comparison of measurements of earnings obtained from a survey with individual responses from employees working for a specific company with the company's own records or with social security records. The company's records are regarded as true measurements. Caution, however, needs to be exercised in generalising the results from such validation studies, since in most cases the research carried out is based on records from a single company to obtain information on earnings and hours worked. The results might therefore not be representative for the whole population. It should also be noted that for a validation study on measurement error one need to be able to obtain 'true' values of the variables of interest. Usually, companies' or tax records are assumed to represent true values and are used for comparisons on observed values. However, these records can also be subject to error. For example there are limitations in the use of tax records since only earnings above a certain level must be reported. In the following we focus on the nature of measurement error in earnings variables.

To be able to investigate the structure of measurement error in earnings variables Bound et al. (1990) use the PSIDVS validation study to compare measurements of earnings from employees with the company's own records. A particular emphasis lies on the measurement of hourly earnings, which is prone to measurement error because of difficulties in the way of measuring hourly pay and because of difficulties in reporting earnings and hours worked. The PSIDVS data include information on annual earnings and hours such that a variable on hourly earnings can be derived and a validation study for such a derived variable can be carried out. However, the data does not include a direct variable, measuring hourly earnings directly, such that an analysis of the quality of such a direct variable can therefore not be conducted using the PSIDVS data. Bound et al. (1990) investigate the quality of a derived hourly earnings variable. They find that the amount of measurement error in reports of annual earnings is low, in annual work hours it is higher and errors in reports of hourly earnings, derived by dividing annual earnings by annual hours, is quite high. Similar results are obtained by Pischke (1995) and Duncan and Hill (1985) who also report a significant amount of measurement error in such a derived variable. Bound et al. (1990) also investigate alternative measurements of hourly earnings, which, however, are not found to improve the measurements on the derived variable based on annual data. Various measurements exist for deriving hourly earnings and three approaches are described in greater detail. The first method is to obtain information on earnings for a specific, normally the most recent pay period.

Another strategy is to ask about earnings referring to a longer period of time, typically one calendar year. It is assumed that this longer period reduces measurement errors due to tax returns etc. However, difficulties may arise if respondents change their jobs, are unemployed for some of that time or simply cannot remember the amount of earnings over such a long time period. A third possibility is to ask about "usual" hours and "usual" earnings. However, none of the strategies seems to improve the estimates for hourly earnings much, with estimates based on annual earnings and annual hours showing the best results. In addition, Moore, Stinson and Welniak (2000) find that annual reports perform better and that in comparison errors in monthly reports seem slightly larger.

1.2.2 Violation of Common Measurement Error Assumptions

In classical measurement error models certain assumptions are made about the error structure, which might not hold in reality. The classical approach describes measurement error as fully random and reflects non-systematic misreporting among the respondents. However, systematic errors are possible. Bound et al. (1990 and 1994) find a mean-reverting measurement error, i.e. $\mathrm{cov}_{DM}(y_i, \delta_i) \neq 0$ (see also Bound and Krueger, 1991), with a pronounced negative correlation between the error and the true level of the variable. In standard measurement error models such correlations are assumed to be zero, but in their study this assumption is clearly violated. Classical models also assume covariances between measurement errors in each measure and between measurement and the true level of other variables to be zero. In their study, however, they find evidence of non-zero covariances. Also Pischke (1995) reports that measurement errors in earnings seem to be correlated with many typical regressors in earnings equations. Bound et al. (1994) investigate the consequences of measurement error on estimates for a regression model, where for example the logarithm of annual earnings is the dependent variable. They find that errors in measuring earnings are related to some of the explanatory variables. Rodgers and Herzog (1987) emphasize that in their study commonly made assumptions about the independence of measurement error are often incorrect. They find substantial correlation between measurement errors and correlations between errors of one variable and measured values on other variables. One approach of reducing the impact of measurement error effects in the data analysis stage is to transform the scale on which a variable is measured, for example by using logarithmic or other transformations. The logarithmic transformation is commonly used to handle the case of a

multiplicative error model (Duncan and Hill, 1985; Bound and Krueger, 1991; Rodgers, Brown and Duncan, 1993; Brownstone and Valletta, 1996).

Modelling the measurement error process can be of interest when correcting for measurement error. Brownstone and Valletta (1996) model the measurement error structure from existing validated data to account for measurement error in standard non-validated data. A key issue is the extent to which measurement error is systematically related to demographic and economic variables. It is therefore of interest to model the measurement error structure and its dependence on certain explanatory variables. The following model is considered

$$X - Y = g(Y) + \mu\beta + \delta\,, \tag{1.26}$$

where g is a function, which allows for a non-linear relationship between measurement error and true earnings, μ is a row-vector of functions of covariates, β is a vector of coefficients and δ is an error term. The dependent variable in this regression is [ln(earnings with error) - ln(true earnings)]. In the classical error approach it holds: $g(Y) = 0$, $\beta = 0$ and δ with mean equal to zero and constant variance. Using this model Brownstone and Valletta (1996) find that measurement error is significantly and negatively related to true earnings. In addition, certain explanatory variables have a non-negligible effect on measurement errors, such as gender, job experience, education, industry section and marital status. Bound et al. (1990 and 1994) report that the probability that a variable, such as employment status, is subject to response error is related to a set of demographic factors. Younger and less educated workers for example are more likely to provide erroneous reports of their employment status. This demonstrates the presence of systematic rather than random measurement error in earnings, which can potentially bias estimates based on variables related to earnings.

1.3 Nonresponse in Surveys

1.3.1 Nonresponse and Missing Data

Correcting for measurement error using validation data, where true values are observed for a subset of the initial sample, can be viewed as a missing data problem. Methods compensating for nonresponse, such as imputation, can then be used to fill in the missing values in the true variable. The required analysis can be carried out based on the imputed true variable rather than the observed values subject to measurement error regarding the measurement error problem as a missing data problem. This is particularly important if classical measurement error assumptions are violated and standard adjustment methods in the presence of measurement error might not be appropriate. We therefore review the missing data literature, common assumptions made about the missing data structure and possible imputation methods to adjust for nonresponse bias.

In sample surveys nonresponse is often a major problem. By nonresponse it is meant that the required data are not obtained for all elements, which are selected for observation. Generally, a distinction is made between *unit nonresponse*, i.e. the failure of a selected sample member to respond, and *item nonresponse* where it is failed to obtain some required information from individual sample members. The case of item nonresponse where information on certain variables is available from an external source for a non-sample member is not considered here. Unit nonresponse occurs if it is not possible to interview certain sample members or if sample members did not want to take part in the survey. Item nonresponse on the other hand occurs if the interviewer fails to ask a question, does not record the answer or the sample member refuses to answer a question or does not know the answer. More on the reasons for nonresponse can be found for example in Lessler and Kalsbeek (1992). There are several ways of dealing with nonresponse problems. For unit nonresponse normally *weighting* methods are applied, whereas for item nonresponse *imputation* methods may be used, which are described in greater detail below. In general, the missing-data structure can be *univariate*, that means that the missing values only occur in a single response variable, or *multivariate* in the sense that missing values occur in more than one variable. In the case of nonresponse one major problem is that we usually do not know how the set of respondents for each variable in the data set is generated. However, knowledge or the

absence of knowledge of the mechanisms that led to the missing data structure is a key element in choosing an appropriate method for handling missing data and for interpreting the results. For the general case the distribution $f(\tilde{I} \mid \tilde{H}, \psi)$, where $\tilde{I}$ is the indicator of response, $\tilde{H}$ a vector of variables of interest and ψ is a vector of parameters of the model, is called the nonresponse mechanism, which is usually unknown. The selection of the sample and the occurrence of nonresponse in the sample can be regarded as a two-phase sampling procedure. Since the selection procedure of the respondents in the sample is unknown, it is necessary to make certain assumptions about the response distribution, which often cannot be verified (Kalton and Kasprzyk, 1982; Kalton, 1983; Schulte Nordholt, 1998; Särndal, Swensson and Wretman, 1992; Lessler and Kalsbeek, 1992, Little and Rubin, 2002).

1.3.2 Missing Data Mechanisms

Several assumptions about the response mechanism are possible. One assumption that is sometimes made when considering imputation methods is that the data are *missing completely at random* (MCAR). That means that the probability of response does neither depend on the response variable Y, subject to nonresponse, nor on the explanatory variables W. In this case the observed values of the response variable form a random subsample of the sampled values of Y (Rubin, 1976; Little, 1986; Little, 1988, Dec.; Little and Rubin, 2002). This mechanism is also called *uniform* nonresponse (Deville and Särndal, 1994; Rao, 2001). Under MCAR the probability of response is $pr(i \in r) = c$, for all $i \in s$, $c > 0$ and r denotes the set of respondents. If the sample is divided into mutually exclusive and exhaustive classes, such that the response mechanism is constant within these classes, the nonresponse mechanism is said to be *uniform within classes*. The following definitions refer to the case discussed in this thesis where only the variable Y is subject to nonresponse and the vector of covariates W is fully observed. For completeness we also give the definitions for the general case with several variables subject to nonresponse. The symbol $\perp$ is used to denote independence.

Definition 1.1 (MCAR)

The missing data mechanism *missing completely at random* (MCAR) is defined as

$$I \perp (Y, W), \tag{1.27}$$

i.e. I is independent of Y and W jointly.

According to the definition of independence this means

$$f(I, Y, W) = f(I) f(Y, W). \tag{1.28}$$

MCAR is therefore equivalent to $f(I \mid Y, W) = f(I)$, since

$$f(I \mid Y, W) = \frac{f(I, Y, W)}{f(Y, W)} = \frac{f(I) f(Y, W)}{f(Y, W)} = f(I).$$

Since $$I \perp (Y, W) \Rightarrow I \perp Y \text{ and } I \perp W, \tag{1.29}$$

MCAR implies $f(I \mid Y) = f(I)$ and $f(I \mid W) = f(I)$. Note that the inverse of (1.29) is not true according to Simpson's paradox (Whittaker, 1990, pp. 30-31, p. 45). Because of implication (1.29) and the fact that $I \perp Y \Leftrightarrow Y \perp I$ it follows

$$f(Y \mid I) = f(Y) \text{ and } f(W \mid I) = f(W). \tag{1.30}$$

More generally, Schafer (1997) defines MCAR as

$$f(\tilde{I} \mid \tilde{H}, \psi) = f(\tilde{I} \mid \tilde{H}_{obs}, \tilde{H}_{mis}, \psi) = f(\tilde{I} \mid \psi), \tag{1.31}$$

where $\tilde{I}$ is the indicator matrix of response, $\tilde{H}_{obs}$ the observed data, $\tilde{H}_{mis}$ the missing data and ψ the vector of unknown parameters of the nonresponse model $f(\tilde{I} \mid \tilde{H}, \psi)$.

If the probability of response does not depend on the response variable but on the explanatory variables, it is said that the data are *missing at random* (MAR). That means that the observed values of Y are a random sample of the sampled values within subclasses defined by values of W, but not necessarily a random subsample of the sampled values (Rubin, 1976; Little, 1988, Dec.; Little and Rubin, 2002). This mechanism is also called *uncorfounded* nonresponse (Deville and Särndal, 1994; Rao, 2001).

Definition 1.2 (MAR)

The missing data mechanism *missing at random* (MAR) is defined as

$$I \perp Y \mid W, \tag{1.32}$$

which means $f(I,Y \mid W) = f(I \mid W) f(Y \mid W)$, according to the definition of conditional independence.

Therefore MAR is equivalent to $f(I \mid Y,W) = f(I \mid W)$, (1.33)

since $f(I \mid Y,W) = \frac{f(I,Y,W)}{f(Y,W)} = \frac{f(I,Y,W)/f(W)}{f(Y,W)/f(W)} = \frac{f(I,Y \mid W)}{f(Y \mid W)} = f(I \mid W)$.

Since $I \perp Y \mid W \Leftrightarrow Y \perp I \mid W$,

MAR is equivalent to $f(Y \mid W,I) = f(Y \mid W)$, (1.34)

or alternatively $f(Y \mid W,I=0) = f(Y \mid W,I=1)$. (1.35)

It can easily be seen that MCAR implies MAR, since $f(I \mid Y,W) = f(I)$ implies $f(I \mid Y) = f(I)$ and therefore $f(I \mid Y,W) = f(I \mid W)$, which is equivalent to the definition of MAR. If the auxiliary variables W are categorical MAR is equivalent to uniform within classes nonresponse.

For the general case Rubin (1976 and 1987) defines MAR as follows. The data is missing at random if the distribution of the missing-data mechanism does not depend on the missing values $\tilde{H}_{mis}$, i.e.

$$f(\tilde{I} \mid \tilde{H},\psi) = f(\tilde{I} \mid \tilde{H}_{obs},\tilde{H}_{mis},\psi) = f(\tilde{I} \mid \tilde{H}_{obs},\psi). \quad (1.36)$$

Note that the definition in (1.36) is more general than (1.33) since it also allows the dependence on the observed values for Y. For the definition of MAR it is required that the covariates W are included among the model covariates. If the probability of response depends on Y and possibly also on W, the data are neither MAR nor MCAR and the missing-data mechanism is said to be nonignorable (Schafer, 1997; Little and Rubin, 2002).

1.3.3 Ignorable and Nonignorable Nonresponse

Two terms that are closely related to the ideas of MCAR and MAR are ignorable and nonignorable nonresponse. If either MCAR or MAR hold we say that the missing-data structure is *ignorable*. In

general, the mechanism leading to missing values cannot be ignored. If the probability that an item y_i is missing depends on the variable Y, and therefore neither MAR nor MCAR hold, the missing-data mechanism is *nonignorable*. This is also called *confounded* nonresponse (Deville and Särndal, 1994; Rao, 2001). Standard analysis that ignores the missing data mechanism would then lead to biased results. Initially, we define the nonresponse to be *nonignorable* if the probability of response is neither MAR nor MCAR. To be nonignorable it is therefore enough to be not MAR, which means

$$f(I \mid Y, W) \neq f(I \mid W) \tag{1.37}$$

or alternatively

$$f(Y \mid W, I) \neq f(Y \mid W) \tag{1.38}$$

or

$$f(Y \mid W, I = 1) \neq f(Y \mid W, I = 0). \tag{1.39}$$

In addition to this intuitive definition, we give a more rigorous definition of ignorable and nonignorable nonresponse (Rubin, 1987; Heitjan, 1994; Schafer, 1997; Little and Rubin, 2002). Let ς be the parameter of the model of the complete data, $f(\tilde{H} \mid \varsigma)$, and ψ the parameter of the nonresponse model, $f(\tilde{I} \mid \tilde{H}, \psi)$. We introduce the following definition of two distinct parameters.

Definition 1.3 (Distinct)

The parameters ς and ψ are *distinct* if the joint parameter space of (ς, ψ), that is the set of all possible values that the parameters ς and ψ can assume, is the product of the parameter space of ς and the parameter space of ψ.

Definition 1.4 (Ignorability)

The missing-data structure is said to be *ignorable*, if the data is missing at random (MAR) and the parameters ς of the data model, $f(\tilde{H} \mid \varsigma)$, and the parameters ψ of the missingness mechanism, $f(\tilde{I} \mid \tilde{H}, \psi)$, are distinct.

This gives a more precise definition of the former intuitive description of ignorability. The above definition of distinctness is given from a frequentist perspective. From a Bayesian perspective

distinctness of two parameters means that any joint prior distribution applied to (ς,ψ) must factor into independent marginal priors for ς and ψ, i.e. $f(\varsigma,\psi)=f(\varsigma)f(\psi)$. In many cases knowing ς gives little information about ψ and vice versa, which makes this definition intuitively clear. Assuming that the nonresponse mechanism is ignorable we do not need to take into account $\tilde{I}$ or ψ for likelihood-based or Bayesian inference about ς, i.e. Bayesian and likelihood inferences that ignore the randomness of the mechanism are valid. The following implications of the condition of ignorability are taken from Schafer (1997) and Little and Rubin (2002). In general, the joint distribution of the observed data can be written as

$$f(\tilde{H}_{obs},\tilde{I}\mid\varsigma,\psi)=\int f(\tilde{H}_{obs},\tilde{H}_{mis}\mid\varsigma)\,f(\tilde{I}\mid\tilde{H}_{obs},\tilde{H}_{mis},\psi)\,d\tilde{H}_{mis}, \tag{1.40}$$

which under MAR simplifies to

$$f(\tilde{H}_{obs},\tilde{I}\mid\varsigma,\psi)=f(\tilde{H}_{obs}\mid\varsigma)f(\tilde{I}\mid\tilde{H}_{obs},\psi)\,. \tag{1.41}$$

The likelihood of the observed data can therefore be factorised into two components, one referring to the parameter of interest ς and the other one to ψ. Under the condition that the two parameters are distinct, likelihood-based inferences about ς are not affected by the parameter ψ or $f(\tilde{I}\mid\tilde{H}_{obs},\psi)$. That means that likelihood-based inferences for ς from $L(\varsigma,\psi\mid\tilde{H}_{obs},\tilde{I})$ are the same as from $L(\varsigma\mid\tilde{H}_{obs})$, and maximum likelihood estimation can be performed ignoring the missing-data mechanism. The likelihood ignoring the missing data mechanism is referred to as the observed-data likelihood:

$$L(\varsigma\mid\tilde{H}_{obs})\propto f(\tilde{H}_{obs}\mid\varsigma)\,. \tag{1.42}$$

From a Bayesian perspective it can be shown that under ignorability all information about ς is summarised in the posterior, which does not take into account the missing-data mechanism (Schafer, 1997), i.e.

$$f(\varsigma\mid\tilde{H}_{obs},\tilde{I})=f(\varsigma\mid\tilde{H}_{obs})\propto L(\varsigma\mid\tilde{H}_{obs})f(\varsigma)\,, \tag{1.43}$$

where $f(\varsigma)$ is the prior distribution for ς. The posterior distribution $f(\varsigma\mid\tilde{H}_{obs})$ is referred to as the observed-data posterior.

1.4 Imputation Methods

Imputation is a method to fill in missing data to produce a complete data set. Usually, imputation makes use of a certain number of auxiliary variables that are statistically related to the variable in which item nonresponse occurs. These auxiliary variables should be available for respondents and nonrespondents (Lessler and Kalsbeek, 1992). A distinction may be made between *deterministic* and *stochastic* (or random) methods. Given a selected sample deterministic methods always produce the same imputed value for units with the same characteristics. Stochastic methods may produce different values.

There are a number of reasons for carrying out imputation. The main reason is to reduce nonresponse bias, which occurs because the distribution of the missing values, assuming it was known, generally differs from the distribution of the observed items. When imputation is used, it is possible to recreate a balanced design such that procedures used for analysing complete data can be applied. However, it can have serious negative impacts if imputed values are treated as real values. To estimate the variance of an estimator subject to imputation adequately often special adjustment methods are necessary to correct for the increase in variability due to imputation. It is also possible to increase the bias by using imputation, e.g. if the relationship between known and unknown variables is poor (Kalton and Kasprzyk, 1982; Kalton, 1983; Särndal, Swensson and Wretman, 1992; Little and Rubin, 2002).

To facilitate our discussion we introduce the notation $Y.$ for the imputed variable of Y, such that

$$y_{\cdot i} = \begin{cases} y_i \text{ for } i \in r \\ \hat{y}_i \text{ for } i \in \bar{r} \end{cases}, \tag{1.44}$$

where $i \in s$ and $\hat{y}_i$ denotes the imputed value for the nonrespondent i. The notation r and $\bar{r}$ refers to the set of respondents and nonrespondents respectively. Let θ denote the parameter of interest in the population, which is a function of the values of Y, $\theta = \theta(y_1, ..., y_N)$, and $\hat{\theta}$ an estimator of θ based on the sample in the case of full response, such that $\hat{\theta} = \hat{\theta}(y_1, ..., y_n)$. Applying imputation in the case of nonresponse we obtain an estimator of the form $\hat{\theta}_{\cdot} = \hat{\theta}_{\cdot}(y_{\cdot 1}, ..., y_{\cdot n})$, called the imputed estimator. This imputed estimator may be biased. However, the bias can be small or negligible (Deville and Särndal, 1994; Lee, Rancourt and Särndal, 2000). In

the following sections common imputation methods are discussed. Multiple imputation is briefly reviewed. Some alternative approaches to handling missing data are addressed.

1.4.1 Common Simple Imputation Methods

There are a number of different approaches to imputation. *Deductive methods* impute a missing value by using logical relations between variables and derive a value for the missing item with high probability. The method of *mean imputation* imputes the overall mean of a numeric variable for each missing item within that variable. If it is a categorical variable the *mode* is usually taken. A variation of this method is to impute a class mean, which involves allocating respondents first into several classes. Then the mean of each class is substituted for the missing values within these classes. Provided the classes were chosen appropriately this method may reduce nonresponse bias. Disadvantages of this procedure are that distributions of survey variables are compressed and relationships between variables are normally distorted (Kalton, 1983; Lessler and Kalsbeek, 1992; ONS, 1996; Little and Rubin, 2002).

Many approaches have been developed that assign the value from a record with an observed item to a record with a missing value on that item. These records are often referred to as *donor* and *recipient* respectively (Kalton, 1983). Such an imputation method is sometimes referred to as a *donor method*. A simple donor method is to impute for each missing item the response of a randomly selected case for the variable of interest. Such a method involves consideration of how the respondent and therefore the donor value should be selected. For example stratification can be applied first and the sampling of donors can be carried out with or without replacement. An advantage is that real values are used for imputation. In the following, different forms of donor imputation methods are discussed.

1.4.2 Cold and Hot Deck Imputation

Cold deck imputation takes the imputed value for each missing item from an external source, for example from an administrative record or from a former survey of the same type. The disadvantage is the lack of comparability between the former and the more up-to date data. Another potential disadvantage is that there might not be a matched record for every

nonrespondent in the external source (Sande, 1982; Lessler and Kalsbeek, 1992). This method was one of the first forms of imputation. However, *hot deck procedures* or donor methods, where the data are taken from current values, are often preferred. Hot deck imputation is very common in practice, especially when dealing with categorical data, since imputed values are taken from the dataset itself and therefore represent actually occurring values. Several approaches exist for selecting the donor value. One form of hot deck imputation, which allows randomisation, is *random hot deck imputation within classes*. The classes are defined as homogenous subsets of the sample, for example formed as a cross-classification of auxiliary variables. After the formulation of imputation classes, missing values within each class are replaced by recorded values in the same class. The selection of donor values within each class can be carried out with or without replacement (Kalton and Kasprzyk, 1982; Little, 1986; Lessler and Kalsbeek, 1992). *Weighted hot deck* imputation is a modification of the hot deck procedure, which takes into account individual selection probabilities for the units in the data set. *Sequential hot deck* imputation involves ordering responding and nonresponding units into a sequence. A missing value of a variable is imputed by the nearest preceding observed value within this variable. As a starting value a donor value is selected, which could be, for example, chosen at random (Kalton and Kasprzyk, 1982; Sande, 1982; Kalton, 1983; ONS, 1996). Another way of assigning the initial donor value is to use a cold deck method, where the missing value is replaced by a value from an external source, for example from a prior survey (Little and Rubin, 2002). If the first value in the sequence is a missing value then the chosen donor value is imputed. If the first value is an observed item than this value becomes the donor value. If the next item is missing than this new donor value is imputed and so on. This has the advantage that if the cases are ordered for example geographically it introduces geographical effects. In this procedure imputation classes may be used. The reason for building these classes is to ensure that consecutive records in each of these classes are as similar as possible with respect to the considered variable. The classes could be defined using cross-classification of auxiliary variables. If a match cannot be found some imputation classes may be combined until a match can be found (Lessler and Kalsbeek, 1992). The method of sequential hot deck imputation is essentially a deterministic method. It is often used in censuses and large scale surveys in which a greater number of missing items occur. In a data set that is large enough it can be ensured that an appropriate number of donor values is available. An advantage of this method is that the imputed values are real values. However, it should be noted that the outcome mainly depends on the order of the file. Also some values might be used several times for imputation if more than one missing

value occurs in a row. This might effect the variance of an estimator resulting in a lower precision of survey estimates. *Hierarchical hot deck* imputation uses a larger number of imputation classes in a hierarchical order. Although hierarchical hot deck may not allow for as many auxiliary variables in the model as in regression imputation it may have the advantage of putting less effort into model building techniques. The matching is done in a hierarchical order. If in the first step, which takes into account many auxiliary variables, a class does not contain any donors, the donor value is found by leaving out less important control variables to allow matching. The system is ordered such that a value can always be found at the lowest level of matching.

1.4.3 Nearest-Neighbour Imputation

Nearest-neighbour imputation, also called *distance function matching*, is a donor method where the donor is selected by minimising a specified 'distance' (Kalton, 1983; Lessler and Kalsbeek, 1992; Rancourt, 1999; Chen and Shao, 2000 and 2001). This method involves defining a suitable distance function or measure, where the distance is a function of the auxiliary variables. The distance from the sample member with the missing value to all other fully recorded members of the sample is calculated. The unit with the smallest distance to the unit of interest is identified and its value is substituted for the missing item according to the variable of concern. One way of defining a distance measure is the Euclidean distance. The easiest way is to consider just one continuous auxiliary variable W and to compute the distance D from all respondents to the unit with the missing item, i.e.

$$D_{ji} = | \mathrm{w}_j - \mathrm{w}_i | , \tag{1.45}$$

where j denotes the unit with the missing item, $j \in \bar{r}$, and $i \in r$. If Y is the variable of interest with the missing item to impute, then the missing item is replaced by the value y_{i^*}, where the respondent $i^* \in r$ is the donor for nonrespondent j if

$$D_{ji^*} = \min_i | \mathrm{w}_j - \mathrm{w}_i | . \tag{1.46}$$

This imputation procedure is not suitable for categorical data and is generally used for numeric data. It is possible that some donors are used several times whereas others might not be used at all. The multiple usage of donors can be restricted to a certain number of times a donor is selected for

imputation. For example the distance in (1.45) can be defined by including a factor such as $1+\lambda t$, where λ is the assigned penalty for each usage and t indicates the number of times the donor has been used. The advantage of nearest neighbour imputation is that actually observed values are used for imputation and that it normally maintains relationships between variables. Chen and Shao (2000) prove that the nearest-neighbour approach, using the Euclidean Distance, estimates distributions correctly. Rancourt, Särndal and Lee (1994) show that it provides unbiased estimates, assuming that a linear relationship exists between Y and W. Chen and Shao (2000) prove that this even holds if almost no assumption is made about the model relating W and Y.

1.4.4 Regression Imputation

Another common method for imputing missing data is regression imputation, described in Kalton (1983), Kalton and Kasprzyk (1982), Lessler and Kalsbeek (1992), ONS (1996) and Little and Rubin (2002). It is distinguished between predictive and random regression. *Predictive regression* imputation, also called *deterministic regression* imputation or *conditional mean* imputation, involves the use of one or more auxiliary variables, of which the values are known for complete units and units with missing values in the variable of interest. A regression model is fitted that relates the variable of interest Y to all auxiliary variables, denoted W. The predicted value for the outcome variable Y is used for imputation of the missing values in Y. In the case that missing values also occur in the auxiliary variables, it is possible to fill in these values by other imputation methods. Usually, linear regression is used for numeric variables, whereas for categorical data logistic regression is preferred. Under *random regression imputation* the imputed value for the variable Y is the predicted value from the regression with a residual term added to the outcome value. This allows for randomisation and reflects uncertainty in the predicted value. This residual can be obtained in different ways. One approach is to assume that the residuals have a normal distribution with zero mean and unknown variance. The error variance can be estimated for example from respondent data. The required residual terms are then generated as draws from a normal distribution. A modification of this approach is to generate the residual terms within certain subclasses, where the variance required for drawing the residual terms from the normal distribution is estimated within these subclasses. Another method of finding residuals is to compute the regression residuals from the complete cases and to select an observed residual at random for each nonrespondent.

The advantage of regression imputation is that it can make use of many categorical and numeric variables. The method performs well for numeric data, especially if the variable of interest is strongly related to auxiliary variables. The difference to some of the other methods is that the imputed value is a predicted value either with or without an added on residual and not an actually observed value. The disadvantage of predictive regression imputation is that it distorts the shape of the distribution of the variable Y and also the correlation between variables, which are not used in the regression model. It might also artificially inflate the statistical association between Y and the set of auxiliary variables. The distortion is particularly disturbing if the tails of the distribution are being studied. For example imputing conditional means for missing income underestimates the percentages of cases in poverty even under MCAR (Kalton, 1983). A random regression model maintains the distribution of the variables and allows for the estimation of distributional quantities (Kalton and Kasprzyk, 1982; Kalton, 1983; Schulte Nordholt, 1998). Also important is the robustness against model misspecification. If the regression model is not a good fit the predictive power of the model might be poor (Little and Rubin, 2002).

Another approach that makes use of the regression model is the method of *predictive mean matching imputation* as described in Little (1988, July) and Heitjan and Little (1991) and Heitjan and Landis (1994). The predictive regression model is carried out and the predicted value of Y for nonrespondent j is compared with the predicted values from the respondents. The distance D_{ji} is as defined in (1.45) based on the predicted values. The value of Y from the respondent whose predicted value is 'closest' to the predicted value from the nonrespondent is imputed for the missing item. We have

$$D_{ji^*} = \min_i | \hat{y}_{reg\,j} - \hat{y}_{reg\,i} | . \tag{1.47}$$

where i^* is the donor value for nonrespondent j and $\hat{y}_{reg\,i}$ denotes the predicted value from the regression model for person i. The imputed value is $\hat{y}_j = y_{i^*}$. Predictive mean matching is essentially a deterministic method. Randomisation can be introduced by defining a set of values that are closest to the predicted value and choosing one value out of that set at random for imputation (Schenker and Taylor, 1996; ONS, 1996; Schulte Nordholt, 1998; Little and Rubin, 2002). The method of predictive mean matching combines elements of regression, nearest-neighbour and hot deck imputation. It is an example of a composite method, where elements of

different imputation methods are combined. It is also assumed to be less sensitive to misspecifications of the underlying model than regression imputation (Schenker and Taylor, 1996).

1.4.5 Repeated Imputation

So far we have discussed single value imputation, where one value is imputed for each missing item. It is also possible to use *repeated* imputation, in the sense that M, $M > 1$, values are assigned for each missing item. There are two reasons for using repeated imputation. One reason is to reduce the random component of the variance of the estimator arising from imputation. This reason is emphasised in the method of fractional imputation (Kalton and Kish, 1984; Fay, 1996), which views the resulting estimator as a weighted estimator with fractional weights $1/M$ for each of the imputed values. Examples of fractional imputation are the use of repeated random hot deck and repeated random predictive mean matching imputation. Note that it only makes sense to apply a random imputation method several times as opposed to a deterministic method.

Another reason for using repeated imputation is simplification of variance estimation in the presence of imputation. The method of *multiple* imputation (MI), as proposed by Rubin (1987), is also a form of repeated imputation in the sense that several values are assigned for each missing item. The idea behind this approach is that the repeated imputed values reflect uncertainty about the true but non-observed values. Single value imputation basically treats the imputed values as known and thus, without special adjustment, it cannot reflect sampling variability under a model of nonresponse or uncertainty about the correct model for nonresponse (Little and Rubin, 2002). Provided the repeated imputation are what Rubin calls *proper* multiple imputation the multiple imputation method based upon the use of only complete-data methods provides a simple method for estimating the variance due to the missing data. This is advantageous since many users and analysts of complex surveys and public-use data sets are not familiar with handling specific missing data problems and are not able to derive specific variance estimation techniques in the presence of imputation. If imputation is carried out by repeating a single imputation method, such as regression or hot deck imputation, it is referred to as *improper* multiple imputation (see also Binder and Sun, 1996), which is the same as fractional imputation. Improper multiple imputation requires special adjustment to the variance of an estimator. In the following proper multiple imputation is

defined. Using the definition in Schafer (1997, p. 105) multiple imputations are said to be *proper* if they are independent realizations of $f(\tilde{H}_{mis} \mid \tilde{H}_{obs})$ the posterior predictive distribution of $\tilde{H}_{mis}$. This posterior predictive distribution of the missing data under some complete-data model and prior can be written as

$$f(\tilde{H}_{mis} \mid \tilde{H}_{obs}) = \int f(\tilde{H}_{mis}, \varsigma \mid \tilde{H}_{obs})\, d\varsigma = \int f(\tilde{H}_{mis} \mid \tilde{H}_{obs}, \varsigma)\, f(\varsigma \mid \tilde{H}_{obs})\, d\varsigma . \qquad (1.48)$$

Proper multiple imputations therefore reflect uncertainty about $\tilde{H}_{mis}$ given the parameters of the complete data model and uncertainty about the unknown model parameters ς. This definition is given from a Bayesian perspective. Rubin (1987 and 1996) defines proper multiple imputation from a frequentist perspective without reference to any specific parametric model. This definition is referred to briefly at the end of this section. Applying proper multiple imputation enables us to use the resulting M complete-data sets for performing standard complete-data analysis, combining the results for a single overall inference. The differences in the M results obtained from the M complete-data sets can be seen as a measure of uncertainty caused by missing data. An advantage of the method is therefore that it is possible to produce complete micro-data files that can be used for a variety of analyses. Markov chain Monte Carlo and especially data augmentation algorithms are the most common methods for generating the missing data simulations, since the simulated values of $\tilde{H}_{mis}$ have $f(\tilde{H}_{mis} \mid \tilde{H}_{obs})$ as their stationary distribution. In this sense multiple imputation is a Markov chain Monte Carlo approach to the analysis of incomplete-data sets (Rubin, 1996; Schafer, 1997; Lipsitz, Zhao and Molenberghs, 1998).

After having obtained M different complete-data sets via multiple imputation, the aim is to combine the results from each of the M complete-data analyses and to produce a single-point estimate of θ, the quantity of interest. According to Rubin's formulae (1987, pp. 76-81; see also Heitjan and Rubin, 1990; Schafer, 1997; Little and Rubin, 2002) let $\hat{G}$ denote a variance estimate associated with $\hat{\theta}$ and $\hat{G}.$ is the formula applied to observed and imputed data. Both $\hat{\theta}.$ and $\hat{G}.$ are calculated separately for each data set based on observed and imputed data. The estimates from the mth data set are denoted

$$\hat{\theta}.^{(m)} = \hat{\theta}(\tilde{H}_{obs}, \tilde{H}_{mis}^{(m)}) \text{ and} \qquad (1.49)$$

$$\hat{G}_{\cdot}^{(m)} = \hat{G}(\tilde{H}_{obs}, \tilde{H}_{mis}^{(m)}) \quad \forall\, m = 1, \ldots, M. \tag{1.50}$$

To obtain a combined multiple imputation point estimate of θ the average of the complete-data point estimates are taken, such that we can write

$$\hat{\theta}_{\cdot} = \frac{1}{M}\sum_{m=1}^{M} \hat{\theta}_{\cdot}^{(m)} . \tag{1.51}$$

To obtain a variance estimate associated with $\hat{\theta}_{\cdot}$ calculate the average of the complete-data variance estimates, called the *within-imputation variance*

$$\bar{G}_{\cdot} = \frac{1}{M}\sum_{m=1}^{M} \hat{G}_{\cdot}^{(m)} , \tag{1.52}$$

and the variance estimate of the complete-data point estimates, defined as the *between-imputation variance*,

$$\hat{B}_{\cdot} = \frac{1}{M-1}\sum_{m=1}^{M} (\hat{\theta}_{\cdot}^{(m)} - \hat{\theta}_{\cdot})^2 . \tag{1.53}$$

Combining both forms of the variance estimates including an adjustment term $(1+1/M)$ for finite M, defines the overall variance estimate associated with $\hat{\theta}_{\cdot}$ as

$$\hat{T}_{\cdot} = \bar{G}_{\cdot} + (1+1/M)\hat{B}_{\cdot} . \tag{1.54}$$

If there is no missing information about θ, i.e. $\hat{B}_{\cdot} = 0$, then $\hat{T}_{\cdot} = \bar{G}_{\cdot}$, meaning that $\bar{G}_{\cdot}$ is already the total variance estimate. A 100(1- α)% interval estimate for θ is given by

$$\hat{\theta}_{\cdot} \pm t_{v,1-\alpha/2}\sqrt{\hat{T}_{\cdot}} , \tag{1.55}$$

where v denotes the degrees of freedom, which are given by the formula (Rubin, 1987, p. 77)

$$v = (M-1)[1+b^{-1}]^2 , \tag{1.56}$$

where b is defined as $b = \dfrac{(1+1/M)\hat{B}_{\cdot}}{\bar{G}_{\cdot}}$. The degrees of freedom v are large if M is large and/or $\hat{B}_{\cdot}$ is small. Under these conditions (1.56) approximates the normal distribution.

The above definition of proper multiple imputation is a Bayesian one, whereas most of this thesis will adopt a frequentist approach. Rubin also defines proper multiple imputation from a frequentist perspective (Rubin, 1987; Rubin, 1996; Schafer, 1997). This definition basically means that $\hat{\theta}.$, $\bar{G}.$ and $\hat{B}.$ lead to approximately valid inferences for the complete-data statistics $\hat{\theta}$ and $\hat{G}$ over repeated realizations of the missing-data mechanism. Three conditions for multiple imputation to be proper need to hold: 1.) as $M \to \infty$ we have approximately $(\hat{\theta}. - \hat{\theta})/\sqrt{\hat{B}.} \sim N(0,1)$ over the distribution of the response indicators $\tilde{I}$ with $\tilde{H}$ held fixed. 2.) as $M \to \infty$, $\bar{G}.$ should be a consistent estimate of $\hat{G}$ with $\tilde{I}$ regarded as random and $\tilde{H}$ as fixed. 3.) The true between-imputation variance, i.e. $\hat{B}.$ for $M \to \infty$, should be stable over repeated samples of the complete data $\tilde{H}$, with variability of lower order than that of $\hat{\theta}$. An important result of proper multiple imputation is the following, which evaluates results based on multiple imputation. If we assume that a.) $\hat{\theta}$ and $\hat{G}$ lead to approximately valid complete-data inference for the estimand θ over repeated realizations of the sampling mechanism, i.e. $\hat{\theta}$ is approximately unbiased for θ and $\hat{G}$ is approximately unbiased for the variance of θ, and b.) the multiple imputations are proper, then inference of the incomplete-data analysis given by

$$(\hat{\theta}. - \theta)/\sqrt{\hat{T}.} \sim N(0,1) \tag{1.57}$$

based on $\hat{\theta}.$, $\bar{G}.$ and $\hat{B}.$ leads to approximately valid inference if M tends to infinity.

1.4.6 Evaluation of Imputation Methods and Alternative Approaches to Handling Missing Data

When using imputation there are several potentially important issues to consider. The main reasons for carrying out imputation is to reduce nonresponse bias and to create a complete data set such that in many circumstances standard statistical software can be used, although with restrictions on the application of variance estimation. A danger of imputation is that imputed values might be treated as true values. Imputed data can be subject to imputation error and the effects on bias and variance estimation often cannot be measured easily. Another effect of imputation is that it might increase the variance of an estimator, such that special adjustment methods for estimating the variance in the presence of imputation need to be applied. Different approaches to estimating the variance of an imputed point estimator will be discussed in greater

detail in chapter 4. Another problem of imputation is that statistical association among variables can be underestimated. Relationships between the variable of interest and the covariates can be retained by using methods such as regression, nearest neighbour or hot deck within classes imputation (Sande, 1982; Kalton, 1983; Little, 1988, July; Lessler and Kalsbeek, 1992). When choosing among imputation procedures the type of analysis that needs to be conducted should be taken into account. In particular, it should be distinguished if the goal is to produce efficient estimates of means, totals, proportions and official aggregated statistics or a complete micro-data file that can be used for a variety of different analyses. Desirable characteristics of imputation methods are small mean square error with a small nonresponse bias and robustness against model misspecification. Other issues are the availability of variance estimation formulae and practical questions concerning implementation and computing time.

In addition to the imputation-based methods described here a wide range of other methods exist for handling missing data and compensating for nonresponse bias, which will be discussed briefly. A commonly used method is weighting, which is usually applied in the case of unit nonresponse but can also be used to compensate for item nonresponse. Weighting procedures differentially weight the complete cases to adjust for nonresponse bias. These procedures incorporate weights, which are inversely proportional to the probability of response. The approach can be regarded as analogous to using the inverse of the sampling probabilities to compensate for differential sample selection, such as in the Horvitz-Thompson estimator (Oh and Scheuren, 1983). In practice, the response probability is not known and needs to be estimated based on information available for respondents and nonrespondents. Little and Rubin (2002) discuss the example of weighting adjustment within classes, where the probability of response is defined for all units in each class. An estimate of this probability for each class can be obtained by the ratio of the number of responding units to the number of all units in the class. The weighting classes can be formed from survey design variables or from sampled items recorded for both respondents and nonrespondents. Another method of forming weights is based on the response propensity. The respondents' weights are set proportional to the inverse of response rates. This method is referred to as *propensity score weighting* (David et al., 1983) and is an extension to weighting class adjustment methods. The propensity scores used to adjust for differential nonresponse can be estimated using for example a logistic regression model. Under correct model specifications, this method removes

nonresponse bias but may lead to estimators with a high variance because respondents with low estimated response propensity receive large nonresponse weights. It should be noted that there is a close relationship between imputation and weighting, since hot deck imputation can be viewed as a form of weighting, where the weights are determined by the number of times a respondent has been used as a donor. More on the relationship between imputation and weighting can be found in Oh and Scheuren (1983), David et al. (1983) and Little (1986). Propensity score weighting as well as the relationship between weighting and imputation is investigated in chapter 5.

Other approaches of handling missing data are model-based procedures which define a model for the partially missing data and basing inference on the likelihood under that model. The parameters of the model are estimated for example using maximum likelihood estimation. Such model-based procedures are discussed in detail in Little and Rubin (2002). The data are assumed to be generated by a model described by a density function $f(\tilde{H},\tilde{I}\,|\,\varsigma,\psi)$ indexed by parameters ς and ψ. As mentioned in section 1.3.3 under the assumption of MAR we have

$$f(\tilde{H}_{obs},\tilde{I}\,|\,\varsigma,\psi)=f(\tilde{H}_{obs}\,|\,\varsigma)f(\tilde{I}\,|\,\tilde{H}_{obs},\psi)\,, \tag{1.58}$$

such that the likelihood of the observed data can be factorised into two components based on the parameter of interest, ς and ψ. If the two parameters are distinct, likelihood-based inferences about ς are not affected by the parameter ψ or $f(\tilde{I}\,|\,\tilde{H}_{obs},\psi)$. It follows that likelihood-based inferences for ς from $L(\varsigma,\psi\,|\,\tilde{H}_{obs},\tilde{I})$ are the same as from $L(\varsigma\,|\,\tilde{H}_{obs})$, and maximum likelihood estimation can be performed ignoring the missing-data mechanism. The likelihood $L(\varsigma\,|\,\tilde{H}_{obs})$ based on the observed data can be a complicated function with no obvious maximum. Under certain missing-data patterns, the likelihood factors into different components, where the parameters of interest are distinct and the components correspond to likelihood functions for complete data problems. Maximization of the overall likelihood requires the maximization of each component. In the case where the missing-data does not have a particular form that allow for factorisation of the likelihood or if factorisation is possible but the parameters are not distinct, iterative methods of computation for situations without explicit maximum likelihood estimates need to be considered. Several iterative algorithms can be defined to obtain maximum likelihood estimates, such as the Newton-Raphson algorithm, which is based on the first and second derivative of $L(\varsigma\,|\,\tilde{H}_{obs})$ (Schafer, 1997). Another iterative method is the Expectation-

Maximization (EM) algorithm, which does not require second derivatives to be calculated (Schafer, 1997; Little and Rubin, 2002). The E-step computes the conditional expectation of the missing data given the observed data and the current estimated parameter $\varsigma^{(d)}$, where $d = 1,...,D$ denotes the iteration. Then the expectations are substituted for the missing data. Note that not necessarily the missing values are substituted in the E-step, but rather the missing portions of the complete-data sufficient statistics, since the focus is on the functions of $\tilde{H}_{mis}$ in the complete-data loglikelihood. The M-step maximizes the expected loglikelihood of the E-step obtaining a new estimate of the parameter $\varsigma^{(d+1)}$. The case of estimating regression coefficients if one or more variables in the regression equation are subject to missing data is discussed in Little and Rubin (2002, section 11.4).

Another possibility for handling nonresponse is to adopt a fully Bayesian modelling approach, where prior distributions are specified for unknown parameters in the model. Bayesian analysis requires specifying a likelihood as the conditional density of the data given the parameters and a prior distribution representing knowledge about the parameters prior to data collection. The posterior density is the basis for all Bayesian inference and summarizes all the information about the parameters of interest. Computing the posterior, however, may be difficult since it may involve high-dimensional numerical integration. Methods such as Markov chain Monte Carlo integration can be used to address this computational problem. Methods such as Gibbs sampling and data augmentation are widely used in the missing data context. Multiple imputation, for example based on the data augmentation algorithm, is easily motivated from a Bayesian perspective, and can be used to impute the missing values and to obtain estimates of the parameters of interest. The use of data augmentation under nonignorable nonresponse will be derived in chapter 6. Other model-based methods will also be discussed in chapter 6, however, they are not the main focus in this thesis. Since the aim is to estimate distribution of earnings a fully model-based approach making assumptions of the distribution of interest is avoided. An advantage of such model-based methods, however, is the availability of estimates of the variance, which take into account incompleteness of the data, for example based on the second derivatives of the loglikelihood or using Rubin's rule to obtain a variance estimate based on multiple imputations.

Chapter 2

Estimation of Pay Distributions from the Labour Force Survey

In April 1999 the first National Minimum Wage (NMW) was introduced in Great Britain, covering all business sectors and regions of the country. The rate was set initially at £3.00 per hour for employees aged 18-21 and at £3.60 for people aged 22 and over. Since then moderate increases and modifications of the level of the NMW have been made. Since October 2003 the rate is set to £3.80 for 18-21 year olds and £4.50 for 22 years and older. Young people between 16 and 17 and those on formal apprenticeships are exempt from the legislation (Low Pay Commission, 2003). The NMW was introduced with the stated aim of improving income of the so-called low paid employees whilst supporting a competitive economy and avoiding negative effects on employment and businesses. It aims to reduce inequality, such as gender pay differences, and to address social exclusion whilst recognizing business realities, especially since smaller firms and certain business sectors are sensitive to increased labour costs. The NMW is part of a much wider policy and must be seen in the context of other governmental reforms, such as the 'New Deal' (Low Pay Commission, 1998).

The introduction of the National Minimum Wage (NMW) raised a lot of interest about the effects of such a minimum wage law. Economists and researchers in the social sciences are concerned with many different aspects of the NMW, for example effects on employment, young workers,

gender differences, pay differentials and effects higher up the earnings distribution, effects on different industry sectors and businesses with higher wage costs as well as regional differences (Low Pay Commission, 1998, 2000; Sunley and Martin, 2000; Dickens and Manning, 2002). Measuring the impact and effects of the NMW and informing policy makers on how to set and change NMW levels, require estimation of hourly pay distributions especially of the lower end. In particular it is of interest to estimate the proportion of low paid employees that earn below the NMW, or to measure the proportion of employees that are paid at the NMW or just above (Low Pay Commission, 2001).

In the following we shall consider the estimation of hourly pay distributions. Let F_Y denote the cumulative distribution function of the variable Y, as defined in section 1.1.2, and $I(.)$ the indicator function with binary outcome, indicating if a condition is true or false. The distribution of hourly pay in the population is defined as

$$F_Y(y) = P(Y \leq y) = \frac{1}{N}\sum_{i \in U} I(y_i \leq y), \tag{2.1}$$

where the variable Y denotes true hourly earnings. The parameter of interest θ is $F_Y(y)$, the proportion of employees earning not more than a specific threshold y assuming one job per employee. Second and further jobs are not taken into account. The aim is to find an approximately unbiased estimator of $F_Y(y)$. Estimates of such a parameter can only be obtained if sufficient and reliable data are available. However, it is difficult to obtain reliable data on hourly earnings since such a variable is often prone to nonsampling errors, such as nonresponse and measurement error, as discussed in chapter 1.

The data that have been used by the Office for National Statistics (ONS) to obtain estimates on low pay are the New Earnings Survey (NES) and the Labour Force Survey (LFS). Both sources have advantages and disadvantages when measuring low pay. In the following we concentrate on estimation methods based on LFS data. The LFS is a large survey of households, which includes information on hours worked and earnings of employees. However, the most common way of measuring hourly pay, dividing gross weekly pay by usual weekly hours, appears to be subject to a considerable amount of measurement error, which is thought to lead to substantial upward bias of estimates of $F_Y(y)$. This variable, dividing weekly pay by weekly hours, will be referred to as the

derived variable. Due to the introduction of the NMW an alternative variable was introduced in the LFS, asking employees directly about their hourly pay. This will be defined as the *direct* variable. This variable appears to give much more accurate information than the derived variable (Skinner, Stuttard, Beissel-Durrant and Jenkins, 2002; ONS, 2000). However, it is subject to a high amount of missing data, because many individuals are not able to report their hourly pay.

One approach of measuring low pay using the data available in the LFS is to correct for measurement error in the derived variable, which requires information about the structure of the measurement error. However, since the direct variable is not obtained on a random subset of the sample, methods based on a two-phase sampling approach, such as described in Luo, Stokes and Sager (1998), cannot be easily applied. It is also not desirable to make simplifying assumptions about the measurement error structure such as that the measurement error follows an additive model and is normally distributed with mean zero and constant variance, since, as shown in chapter 6, such assumptions are likely to be violated. To estimate the distribution of hourly pay it was therefore decided to concentrate on the direct variable, regarding the measurement error problem as a missing data problem. Missing values of the direct variable are imputed by ONS taking into account information on the erroneous variable and other covariates (Stuttard and Jenkins, 2001). The imputation approach is intended to compensate for potential distorting effects of measurement error in hourly earnings and uses the derived variable and other covariates to impute the missing values in the direct variable. The problem of correcting for measurement error is therefore approached as a missing data problem.

This chapter is structured as follows. In section 2.1 the two main data sources about hourly earnings, the NES and the LFS, are described. The focus is on LFS data, in particular on variables available in the LFS for measuring hourly earnings. Section 2.2 presents the initial imputation method used by the ONS to produce estimates on low pay. Estimates of the proportion of low paid employees and estimated pay distributions based on the proposed imputation method as well as simpler estimation approaches are presented.

It should be noted that there are certain employees that are exempt from the NMW legislation, for example employees that receive training within the first six months of their employment, apprentices etc. The estimates about the low paid that earn below the NMW threshold can

therefore not necessarily be used as a measure of non-compliance with the legislation, since neither the NES nor the LFS includes an indicator if a person is eligible for NMW rates or not.

2.1 The Data and Variables Measuring Hourly Pay

2.1.1 Comparison of NES and LFS Data

Regarding information on earnings there are two main sources in Great Britain, which both have been used by the ONS to produce estimates on low pay. The first source is the *New Earnings Survey* (NES), an employer's survey, which is a sample of 1% of employees. The survey is conducted annually and has been held in April each year since 1970. Since 1979 a panel of employees has been selected on the basis of National Insurance numbers. It is a large survey including about 150,000 employees. In the NES earnings data are provided by employers from payroll records and therefore the information is assumed to have a high level of accuracy. The NES provides a detailed breakdown of earnings into its main components such as basic pay, overtime etc. Weekly average hours of work paid at the basic rate and the number of hours worked overtime are given for a certain pay period. This for example allows for the derivation of gross weekly and hourly earnings based on the basic rate only or including overtime, additions to basic pay etc. This is important since the NMW legislation is based on hourly earnings excluding overtime. The main limitation of the NES is that it is based on Pay-as-You-Earn (PAYE) records so that employees who earn less than the weekly PAYE threshold are not included in the sample. This means that low paid employees are under-represented in the NES and that NES estimates therefore underestimate the number of people that are low paid. The NES does not have a weighting system such that population estimates are biased if under-representation occurs.

The other source is the *Labour Force Survey* (LFS), which is a survey of about 60,000 nationally representative households, conducted quarterly since 1992, representing about 0.3% of the population in Great Britain. In comparison to the NES, the LFS has a better coverage of jobs especially low paid jobs. However, this survey contains only about 17,000 employees per quarter and is therefore much smaller than the NES. The LFS is mostly based on telephone interviews and includes proxy response from other members of the household. It is therefore subject to

greater measurement error. The LFS collects a wide range of information about employment, like type of employment, industry, occupation, and details about main and second jobs. In the LFS, however, less detailed information about earnings than in the NES is given. Only gross pay the last time a person was paid is recorded. Much more information is provided in terms of hours worked, which are broken down into basic hours, hours overtime, actual hours worked in the reference week and usual weekly hours worked. Derived hourly earnings in the LFS normally include paid overtime (ONS, 1999, LFS User Guide).

Since both surveys have advantages and disadvantages in the way they collect data on earnings it is often not possible to derive earnings estimates from the LFS that are directly comparable with NES measures. Comparing both datasets shows that, for example, average earnings estimates from the LFS are consistently lower than estimates from the NES and that estimates for the proportion of the low paid are consistently higher from the LFS than from the NES. Because of the problems of coverage of the low paid it is assumed that the LFS provides a more accurate picture of the extent of low pay in Great Britain. In the following we concentrate on the LFS and the methods used to derive estimates of low pay based on LFS data. Estimates based on NES data have also been improved by introducing a weighting system. However, this will not be discussed here.

2.1.2 Sampling Design and Estimation of the Labour Force Survey

The LFS is conducted quarterly each year in Spring (March-May), Summer (June-August), Autumn (September-November) and Winter (December-February), which is done since many labour markets show seasonal variations. For interviewing purposes Great Britain is stratified into 110 Interviewer Areas (IA), which make up the strata. Within each stratum a systematic sampling procedure of addresses is carried out where a sample of addresses with a random start and constant interval is drawn from the Postcode Address File (PAF), such that the addresses are the primary sampling units (PSU's). All adults of a selected household are included in the sample. Each address has an equal probability of selection and is not based on a multistage sampling procedure. The total sample size each quarter is about 60,000 households. This is made up of five rotation groups each of approximately 12,000 private households. For each quarter one of the five groups or subsamples is 'rotated out' and replaced by a newly selected subsample, such that 80%

of selected households are retained in the sample in successive quarters. Each household therefore remains in the sample for five quarters and has five interviews. The interviews are held face-to-face and by telephone. Each of the interviewer areas is divided into 13 'stint' areas, such that each interviewer can interview one stint area per week. These 13 areas within an IA are allocated randomly to the 13 weeks of one quarter for interviewing. In the subsequent four quarters the selected addresses are interviewed in the same week of the quarter. The LFS is a panel survey of addresses and households, which refuse further participation, are not revisited at the next quarter but remain part of the eligible sample (ONS, 1999, LFS User Guide). The LFS provides information on earnings, which is collected since September 1997 in the first and fifth interview of respondents and is therefore based on 24,000 households, generating a sample of about 17,000 employees per quarter.

To produce population estimates and to compensate for differential nonresponse among different sub groups in the population survey weights are constructed. Each case is given a weight, which can be thought of as the number of people that case represents in the population. These weights are constructed by an iterative poststratification procedure (raking), which requires certain known population control totals. The subgroups that make up the poststrata, mainly depend on region, sex and age groups. Earnings data is weighted separately. A small number of variables, which are likely to be important determinants of income are chosen for the weighting process, i.e. sex, age, region, occupation, industry and whether an employee works full or part-time (ONS, 1999, LFS User Giude; Holmes and Skinner, 2000). It should be noted that the existence of two different weighting schemes may lead to inconsistencies between estimates. However, here we will only concentrate on the survey weights for earnings data. More information on the survey weights is given in section 3.2.

2.1.3 Variables Measuring Hourly Pay in the Labour Force Survey

In the following the term 'hourly pay' or 'hourly earnings' may generally be interpreted as how much a person earns per hour. There are two variables that measure hourly pay. The term *derived hourly pay* refers to the derived variable in the LFS, denoted X, where gross weekly earnings are divided by basic usual weekly hours. If an individual usually works overtime then usual paid

overtime hours are included. It should be noted that gross weekly earnings itself is a derived variable dividing gross pay by the pay period that this covered.

The other possibility of obtaining information about hourly earnings is to ask people directly about their hourly earnings. This *direct variable*, denoted Y, refers to the question in the LFS "What is your (basic) hourly rate?". This question was initially only addressed to a subset of all employees in the sample, namely to employees whose pay period is less than monthly or who get their pay as a lump sum or do not know their pay period. This refers to about 25% of all employees in the sample for the March-May 1999 quarter. Since March-May 2000 the question was addressed to all individuals in the sample that answered 'yes' to the previously asked question: "Are you paid a fixed hourly rate?". This increased the total percentage of valid answers to the direct variable to approximately 43%, since many employees are not paid hourly or do not know their hourly rate (ONS, 1999, Oct.). Note that the response rate to the direct variable is as expected higher for lower paid employees. For example, the response rate is approximately 72% for those in the bottom decile of the derived variable. The variable is therefore subject to a high amount of missing data and is only available for a non-random subsample of the LFS, since hourly paid employees are not representative for the whole population.

These two variables play a crucial role when deriving information about hourly earnings. However, the two variables do not lead to the same estimates of earnings and have strengths and limitations when deriving estimates in practice, which are discussed in the following. There is concern that the derived hourly earnings variable may be subject to a considerable amount of measurement error. Experiences from other surveys including variables on earnings and hours have shown that such a derived hourly earnings variable is strongly affected by measurement error (Bound et al., 1990; Pischke, 1995; Duncan and Hill 1985). Responses to the direct variable, however, seem to provide more accurate information. In the following evidence of measurement error and differences in the distributions of the two variables are described.

The (weighted) cumulative distributions of the direct and the derived variable for respondents for whom both variables are observed for the June-August 1999 quarter are shown in figure 2.1. It can be seen that the two distributions differ, in particular in the lower end. Below the first quartile of the two distributions the percentiles for the direct variable are higher than for the derived variable. This is particularly apparent below the first decile. The graph also shows a truncation effect in the

direct variable for the lower end of the distribution, reflecting the level of the NMW, which cannot be observed in the distribution of the derived variable. Above the third decile of the two distributions the relationship of the two variables is reversed and the percentiles for the direct variable are lower than for the derived variable. Also the variance of the derived variable is with 9.28 higher than for the direct variable with 7.46. These differences are thought to be caused by measurement error in the derived variable, leading to a higher dispersion in this variable. The mean of the two variables are very close with £5.3 per hour for the direct variable and £5.5 for the derived variable taking into account respondents to both variables only. However, taking into account all respondents to each of the two variables, the mean of the direct variable is with £5.3 considerably lower than the mean of derived hourly pay with £8.19. This is an indication that the direct variable mostly covers people that are less well paid, since people with a fixed hourly rate are not typical for the whole population and are more likely to earn less.

In addition, we consider the relationship between the direct and the derived hourly pay variable. Figure 2.2 shows the scatterplot of the derived variable and the direct variable based on respondents aged 22+ for the quarter March-May 2000. The two lines in the plot indicate the level of the NMW of £3.60 per hour. The scatterplot gives an indication of a linear relationship between these two variables. However, many discrepancies between the two variables can be found, indicating the presence of measurement error. The direct variable shows clear evidence of a truncation effect at the NMW level of £3.60 per hour for employees aged 22+, which would be expected. Similar 'step' effects can be found in the direct variable at other thresholds such as £10, £13 and £15 per hour. These effects cannot be seen in the derived variable, neither at the NMW level nor at other thresholds. In particular, the derived variable shows a large proportion of employees being paid below the NMW, including employees with unreasonably low hourly earnings between £0 and £2 per hour, indicating the presence of measurement error. It should be noted that also for other LFS quarters it was found that the direct variable shows a significant effect on the earnings distribution at the time when the NMW was introduced, whereas derived hourly pay is not sensitive to this change. Also later changes in the level of the NMW are reflected in the distribution of the direct variable Y but only marginally in the derived variable X.

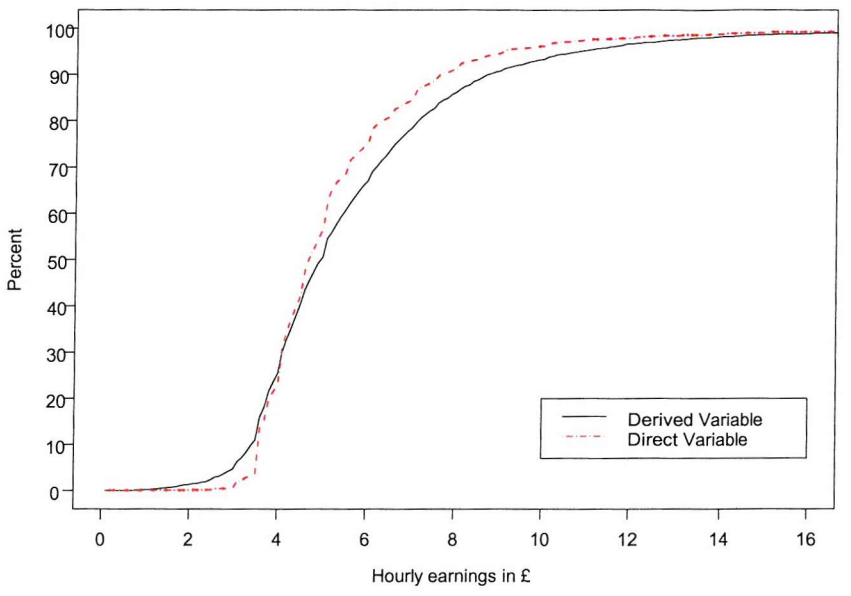

Figure 2.1: Cumulative distributions (weighted) of the direct and the derived variable for cases where both variables are observed, LFS June-August 1999.

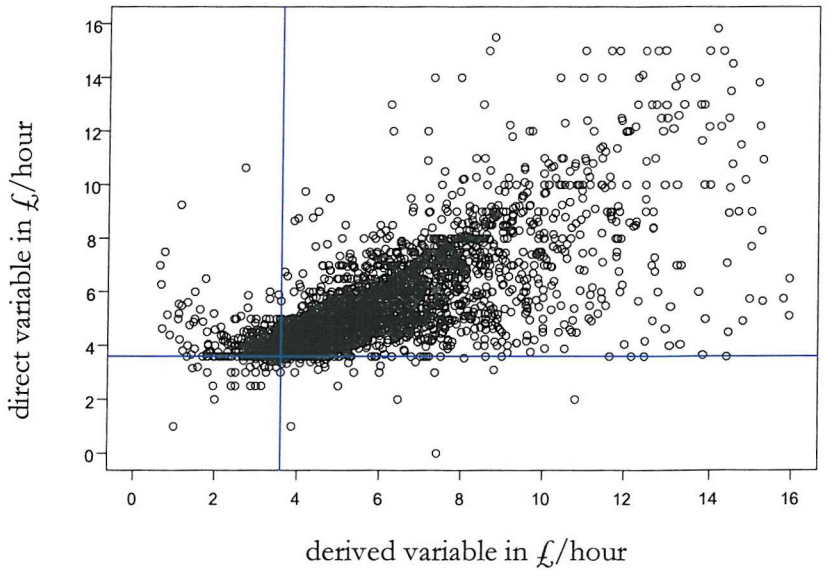

Figure 2.2: Scatterplot of derived hourly pay and the direct variable for the quarter March-May 2000 for cases where both variables are observed, for 22+ age group.

There are several reasons why the distributions for the two variables do not necessarily give the same results. In particular, we discuss the causes that lead to greater measurement error in the derived variable. In the following two types of measurement error are distinguished, referring to definitional error and response error. Definitional error occurs if the respondent answers all questions involved in deriving the 'derived hourly pay' variable 'correctly', however, the value of hourly pay calculated based on these answers is not equal to true hourly earnings. Response error occurs if a person gives inaccurate or incorrect answers.

There are several sources for definitional error in the derived variable. First, hours worked and pay received may refer to different time periods, such that the information derived will be misleading. In particular the information on earnings in the numerator refers to actual earnings whereas the information on hours refers to usual hours. If usual hours and actual hours differ for the reference period measurement error is introduced. Also, the derived variable includes not just the basic pay but also overtime etc. However, the NMW refers to basic pay, not including overtime. In addition, the variable in the numerator 'gross weekly earnings' is a derived variable itself and this variable may already be subject to measurement error. Derived hourly pay is based on the last time a person was paid, which could be up to several weeks ago. This variable, therefore, might not refer to a current rate and might not reflect recent changes in pay. This is an important factor especially around the time when the NMW was introduced and each time the minimum rate is increased. In addition to definitional error, there are reasons to believe that the derived variable is subject to response error. Approximately 30% of responses are received from proxies, which is assumed to lead to inaccuracy in the response. Other response error such as rounding and truncation are also likely to occur. Particularly the information given on hours worked are expected to be approximated since the question asked refers to usual hours worked.

In comparison, the direct variable is assumed to lead to more accurate information of hourly earnings for several reasons. Due to the way the information is obtained the direct variable is not subject to definitional error as the derived variable. In particular, the direct variable refers to basic pay and does not include overtime, which is the measure required by the NMW legislation. It also has the advantage that it refers to a current rate, not the one that applied the last time a person was paid and not necessarily one that applies to the reference week. That means that the direct variable is more up-to-date than the derived variable and covers recent changes in wages. This is important for the time of introduction of the NMW and for changes in NMW levels. The direct variable may

still be subject to response error, such as proxy response, rounding and truncation effects and other sources of response error. Results from the LFS pilot study, however, suggest that the variable works well and is well understood by respondents. In particular, it was found that proxy respondents may respond to the question "What is your (basic) hourly rate?" but do not tend to respond to the direct variable if they do not know the hourly pay rate. This leads to more accurate information, but also to higher nonresponse rates.

A more detailed discussion of the nature and extent of the measurement error particularly in the derived variable is provided in ONS (2000), Stuttart and Jenkins (2001), Skinner and Beissel (2001), Skinner et al. (2002). We conclude that the direct variable appears to give more accurate information on hourly earnings with respect to low paid employees than the derived variable. It is assumed that the derived variable might overestimate the number of employees earning below the NMW and may not be a sensitive indicator of the impact of the NMW. Based on this evidence, the following assumption is made in this thesis.

Assumption 2.1

The direct variable Y reflects true hourly earnings, whereas the derived variable X is subject to measurement error.

This assumption is unlikely to be exactly true but it provides a working assumption which may be expected to lead to more accurate estimates than those produced using the derived variable alone. The key problem with this variable is that it is only available for a subset of the whole sample, i.e. it is subject to a considerable amount of missing data. One possibility to handle missing data is to use imputation. Such an imputation procedure has been implemented by ONS and is described in the following section.

2.2 Distribution of Interest and Initial ONS Imputation Method

This section describes the imputation method, which was initially used by the ONS. It is a random hot deck imputation within imputation classes. Regression imputation is also considered briefly. The underlying aim of this imputation approach is to estimate the distribution of hourly pay as

defined in (2.1), in particular the lower part of this distribution, based on the direct variable and to essentially correct for the measurement error in the derived variable of hourly earnings.

2.2.1 Framework

Before discussing appropriate imputation methods it is necessary to clarify which data are observed and to introduce assumptions about missingness. The direct variable Y is observed for approximately 43% of employees in the sample (March-May 2000 quarter). Only very few cases have missing values in either the derived variable X or in one or more of the covariates W. Therefore nonresponse in the derived variable and other covariates is omitted. We can estimate the following distributions,

$$f(Y \mid I=1),\ f(X \mid I=1),\ f(X \mid I=0),\ f(W \mid I=1) \text{ and } f(W \mid I=0), \tag{2.2}$$

where I indicates if Y is observed or missing. However, we do not observe $f(Y \mid I=0)$. The objective is to estimate $f(Y \mid I=0)$ and to impute the missing values of the variable Y. Since this distribution is unknown certain assumptions are necessary and the aim is to make these assumptions as realistic as possible. A naive approach would be to assume that the observed and the missing values have the same distribution, i.e.

$$f(Y \mid I=0) = f(Y \mid I=1), \tag{2.3}$$

which means that Y and I are independent. Under this assumption the nonresponse would follow the assumption of MCAR (definition 1.1). However, it is unlikely that this assumption holds in reality, since employees that are paid hourly and state their hourly rate are not representative of all employees. Those who report the direct variable are much more likely to be low paid than employees who are not paid on an hourly basis but for example monthly. Investigating the distribution of the derived hourly pay variable X for I=1 and I=0 separately indicates that

$$f(X \mid I=0) \neq f(X \mid I=1). \tag{2.4}$$

For example for the March-May 2000 quarter of the LFS the mean of the derived variable for the cases where the direct variable is not observed is £10.47 and for the cases where it is observed it is

£6.45 (weighted), which shows that the mean for employees that responded, and are therefore paid on an hourly basis, is considerably lower than for those who did not respond. This implies that I is not independent of X such that the assumption of MCAR does not hold. It also suggests that probably $f(Y \mid I=0) \neq f(Y \mid I=1)$. Another simple estimation approach would be to use the values for Y where it is observed and the values of the derived variable where Y is missing. However, it is expected that this approach also leads to upward bias in the estimates of the lower end of the pay distribution. The effect can be seen in figure 2.5. We therefore need to make other more realistic assumptions for estimating the distribution of the missing values of Y.

A more realistic assumption than MCAR, which is commonly used in the imputation literature (Rubin, 1987) is that Y is missing at random (MAR) conditional on the variables X and W observed for all units in the sample (see definition 1.2).

Assumption 2.2 (MAR)

Y is conditionally independent of I given the covariates X and W, i.e. $Y \perp I \mid X, W$.

It follows $$f(Y \mid X, W, I=0) = f(Y \mid X, W, I=1) = f(Y \mid X, W). \qquad (2.5)$$

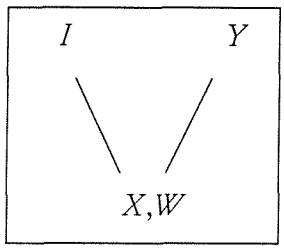

Figure 2.3: Graph of conditional independence under MAR.

We can also write $I \perp Y \mid X, W$ and therefore $f(I \mid Y, X, W) = f(I \mid X, W)$. This means that the probability of response, i.e. the distribution of I, only depends on the covariates X and W but not on Y itself. The graph of conditional independence under the MAR assumption is shown in figure 2.3. Although in reality it is likely that I depends on Y unconditionally, many variables related to pay are observed in the LFS, in particular the derived variable X, and it may be argued that conditioning on such a set of auxiliary variables will reduce or even wipe out the dependence of I

on Y. In the following sections variations of classical regression imputation are presented based on the assumption of MAR. An alternative identifying assumption, under which MAR need not hold, as well as possible adjustment methods based on this assumption are discussed in chapter 6.

2.2.2 Regression Imputation

When applying an imputation method we ideally would like to generate imputed values from the conditional distribution of true hourly pay Y given the derived variable X and other covariates W. This condition can be expressed as

$$f(Y.|X,W,I=0)=f(Y|X,W,I=0), \tag{2.6}$$

where $Y.$ is defined in (1.44). Condition (2.6) means that the distribution of the imputed values given X and W for the nonrespondents is the same as the conditional distribution of the true values. Condition (2.6) is a desirable property of an imputation method since if this condition holds an imputed estimator, i.e. an estimator subject to observed and imputed values of Y, will have the same properties as an estimator based on the true variable Y. The imputed estimator will then be unbiased for the population parameter of interest. However, the true values of Y are not observed for the nonrespondents such that the distribution $f(Y|X,W,I=0)$ cannot be fitted directly. An identifying assumption is therefore required. Under the MAR assumption, which implies $f(Y|X,W,I=0)=f(Y|X,W,I=1)$, it follows for the generation of values for Y from the covariates X and W for I=0, that

$$f(Y.|X,W,I=0)=f(Y|X,W,I=0)=f(Y|X,W,I=1). \tag{2.7}$$

Assuming MAR the aim is therefore to apply an imputation method that enables the generation of imputed values from the conditional distribution of Y given the values of the derived variable and the covariates based on the respondents. In the following we first discuss a parametric approach to imputation involving a regression model such that (2.7) is fulfilled. For the LFS application considered here we shall consider a model of the form

$$\ln(y_i)=\eta_i\beta+\varepsilon_i, \tag{2.8}$$

where η is a row-vector of covariates, functions of the derived variable X and other covariates W, and β is a vector of coefficients. The dependent variable is the logarithm of the direct variable. Also for the derived variable X the logarithmic transformation is used. The use of the logarithm is common in earnings variables to approximate normality (Brownstone and Valletta, 1996; Bound et al. 1994). Note that the linear relationship between the derived and direct variable is more pronounced when plotted on the logarithmic scale as shown in figure 2.4 in comparison to figure 2.2, suggesting the use of log-transformations in further analyses.

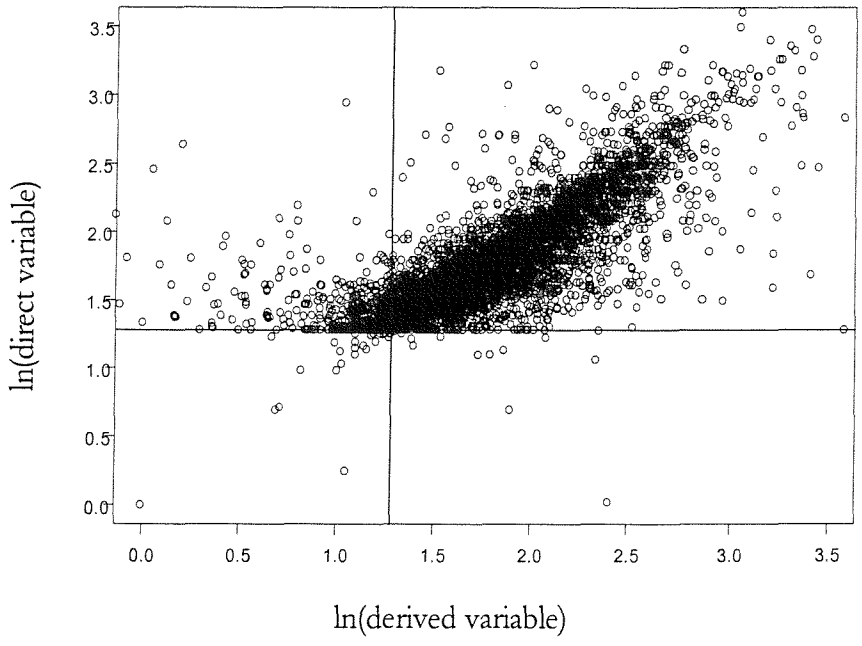

Figure 2.4: Scatterplot of ln(derived hourly pay) and ln(direct variable) for the quarter March-May 2000 for cases where both variables are observed, for 22+ age group.

The regression model (2.8) is described in greater detail in section 2.2.4, and includes predictors of hourly pay and variables that relate to the measurement error. It follows from (2.7) that estimates for the parameters may be based on respondents only. Applying the least squares method, we have

$$\ln(y_i) = \eta_i \hat{\beta} + \hat{\varepsilon}_i \qquad \forall i \in r. \tag{2.9}$$

A predicted value for $\ln(Y)$ for all individuals in the sample is obtained, such that

$$\ln(\hat{y}_{regi}) = \eta_i \hat{\beta} \qquad \forall i \in s. \tag{2.10}$$

These mean values are the base of the following imputation methods. One simple approach for imputation is to use the mean values from the regression as the imputed values. This would, however, artificially reduce the variation in the estimated distribution of interest (Little, 1988; Little and Rubin, 2002). To obtain unbiased predictions one would need to consider the following relationship. If $E[\ln(y_i)] = \eta_i \beta$ it follows that $E[y_i] = \exp(\eta_i \beta + \sigma_\varepsilon^2 / 2)$, where σ_ε^2 is the variance of the residuals, by properties of the lognormal distribution. The usual adjustment for the mean value for Y is thus $\hat{y}_i = \exp(\eta_i \hat{\beta} + s_\varepsilon^2 / 2)$ (see David et al. 1986). However, since the mean values are not used for the actual imputation but only for the definition of classes, as explained in section 2.2.3, we do not need to consider this adjustment. To preserve the distribution noise needs to be added to the mean values at log-scale, which can be done, for example, by adding on random residuals to the mean values drawn from a distribution, such that

$$\ln(\hat{y}_j) = \eta_j \hat{\beta} + \hat{\varepsilon}_j^* \qquad \forall j \in \bar{r}, \tag{2.11}$$

where $\hat{\varepsilon}_j^* \sim N(0, \hat{\sigma}_\varepsilon^2)$ and $\hat{\sigma}_\varepsilon^2$ is the usual least squares estimate based upon respondent data. However, this may perform poorly as shown in David et al. (1986). Another approach of random regression to preserve the distribution is to add on observed residuals, i.e. adding on residuals selected from the set of the observed residuals $\hat{\varepsilon}_i$, where $i \in r$. The method is described in detail in David et al. (1986) (see also Little, 1988; Laaksonen, 1991). The selection of residuals can be carried out within classes and the probability of selection of a residual from respondent i may depend on the survey weight of that respondent i, which will be described in greater detail in section 2.2.3. This imputation method of adding on observed error is less dependent on the assumption of normality and constant variance. The method of random regression imputation adding on random or observed residuals fails, however, to take account of limits on the distribution of the variable of interest. The level of the NMW is such a limit and its existence can be observed in the distribution of the direct variable (see for example figures 2.1 and 2.2). In addition, pay is often rounded, such that step increases throughout the pay distribution can be observed, which leads to a non-smooth distribution of hourly pay. The parametric approaches,

however, do not capture these features. To address this problem semi-parametric approaches can be used. Examples are hot deck imputation methods such as predictive mean matching imputation (Little, 1988).

2.2.3 Hot Deck Imputation Within Classes

The following imputation method has been used by ONS. The imputation method requires the above regression model (2.8) to be estimated and the predicted values $\hat{y}_{regi}$ to be computed for respondents and nonrespondents. Based on successive intervals of values of $\hat{y}_{regi}$, K imputation classes are formed, denoted B_k, where $k = 1,...,K$. Within each imputation class, the imputed value, denoted $\hat{y}_j$, for an employee j without a value for Y, called the *recipient*, is obtained from the value for a person i in the same imputation class with an observed value for Y, called the *donor*. Hence

$$\hat{y}_j \in \{ y_i \mid i \in B_{k(j)} \cap r \}, \quad j \in \bar{r} \tag{2.12}$$

where $\bar{r}$ denotes the set of nonrespondents, r the set of respondents and $B_{k(j)}$ the imputation class to which recipient j belongs. The selection of donors in the imputation procedure may depend on the survey weights in the following way. Respondent i, $i \in B_k \cap r$, is selected as a donor for nonrespondent $j \in B_k \cap \bar{r}$, with probability $w_i / \sum_{i \in B_k \cap r} w_i$ (see also Rao and Shao, 1992). The selection can be carried out with or without replacement. The without replacement sampling has the advantage that it reduces imputation variability and avoids the problem that some donors are used disproportionately often (Kalton, 1983). In the following chapters also the case of selecting donor values with replacement to obtain independent imputations is considered.

Since the above imputation method is random, additional variation is introduced into the estimation process, which could have a nonnegligible impact on the overall variance of the estimated distribution especially if the nonresponse rate is high. Thus, it was decided to repeat the imputation method, such that M values are imputed for each missing observation (in the ONS methodology M=10). This way the impact of the random component in the variance can be reduced, especially if M is reasonably large. In the following we refer to this method as *repeated* imputation. Note that the method used here does not reflect *proper* multiple imputation in the

sense of Rubin (1987; see also section 1.4.5). Such a form of repeated imputation method is also called *fractional* imputation (Kalton and Kish, 1984; Fay, 1996), since each imputed value is allocated the weight $1/M$. We obtain M variables, denoted $Y_{\cdot m}$, where

$$y_{\cdot mi} = \begin{cases} y_i & \text{for } i \in r \\ \hat{y}_{mi} & \text{for } i \in \bar{r} \end{cases} \quad \text{for } m=1,\ldots,M, \tag{2.13}$$

and $\hat{y}_{mi}$ denotes the mth imputed value for nonrespondent i. This semi-parametric imputation method is a random hot deck imputation, which makes use of the regression predictor but is not fully reliant on model assumptions such as the previously described regression methods. It should be noted that, since the assumption of MAR might not hold exactly but is only an approximation, the imputation method is best applicable to estimation of the lower end of the earnings distribution where the proportion of response is largest. An advantage of such a hot deck method is that only eligible values of the missing variable are imputed, which helps to reproduce the shape of the distribution (Rao and Shao, 1992; Little, 1988; Laaksonen, 1992). Limits such as the NMW are respected and heaping at round values of pay rates should be preserved. Hot deck imputations are conservative, in the sense that they do not extrapolate nonrespondent values outside the range of the respondent data. Using hot deck imputation instead of regression imputation also allows for non-constant variance in the imputed values. There is also evidence that such a method is less sensitive to model misspecification than regression imputation (Schenker and Taylor, 1996).

The above imputation method can be viewed as a form of predictive mean matching as described in Little (1988). Each nonrespondent is matched to the respondent with the closest predicted mean, where all values for $\hat{y}_{regi}$ within the same imputation class are viewed as equally close. Then the respondent's value is imputed for the missing item, i.e. $\hat{y}_j = y_i$ for $i \in r$ and $j \in \bar{r}$. Various forms and extensions of this method will be discussed in greater detail in chapter 5.

The use of imputation classes or adjustment cells is common in practice. For example the CPS (Current Population Survey) carried out by the U.S. Bureau of the Census uses hot deck imputation within adjustment cells for missing income items (Little, 1988; David et al. 1986). The adjustment cells are formed by putting people into groups according to values of some explanatory variables. David et al. (1986) compare the hot deck procedure with an alternative regression imputation and describe advantages and disadvantages. With the CPS hot deck

imputation the matrix of adjustment cells might be big and cells might need to be merged to find a matching respondent. The advantage of a regression imputation is that more explanatory variables can be incorporated. David et al. evaluate both approaches, the hot deck and the regression imputation, for the CPS using validation data. They conclude that the hot deck method performs well and does not underestimate income aggregates. Using regression imputation adding on random residuals did not perform well in practice. Also regression imputation adding on residuals from respondent did not perform well producing too many large wage and salary amounts. However, using these methods within imputation classes showed much better results (David et al., 1986). The impact of the choice of imputation classes in the application discussed here is investigated in chapter 3.

2.2.4 The Regression Model

The regression model underlying the imputation method represents the conditional distribution of the variable Y given a number of explanatory variables, including the derived variable X. The choice of explanatory variables is important to ensure that the assumption of MAR holds approximately. Therefore different types of covariates need to be considered, including variables that predict I, the indicator of response, and variables that relate to Y. Using X to predict Y we also need to consider additional covariates that may predict measurement error in X, i.e. the discrepancy between X and Y. In the regression model the dependent variable is $\ln(Y)$. The model includes fifteen continuous and categorical explanatory variables, which are shown in table 2.1. It includes two quadratic terms and a number of two-way interactions of covariates with the derived hourly pay variable. Also for the derived variable the logarithmic transformation was used. It was found that the derived variable was the most significant variable in the regression equation. Other variables relate to the type of job, occupation, industry sector, region or ways of payment. A table of coefficients and diagnostic plots are given in the appendix A2.1. Figure A2.1.1 shows the scatterplot of the residuals versus fitted values indicating a random spread with no obvious pattern. However, further analysis described in chapter 6 may indicate departure from the assumption of constant variance. The histogram of the residuals in figure A2.1.2 and the normal probability plot in figure A2.1.3 indicates an approximately normal distribution of the residuals. The model diagnostics will be discussed further in chapter 6. Similar regression models are used in

the CPS imputation method, where the dependent variable is also the logarithm of wages and salaries. Similar explanatory variables are incorporated, such as region, age, education, gender, occupation etc. In addition non-linear effects are included such as age squared and a number of interactions (David et al., 1986).

Name of Variable	Abbreviation	Continuous/ Categorical
Ln of derived hourly pay $= \ln(X)$	LHE	continuous
Squared ln of derived hourly pay $= \ln^2(X)$	LHE^2	continuous
Major occupation group	SOCcat	9 categories
Part-time/ Full-time	PT	2 categories
Paid less than weekly	LTWK	2 categories
Highest qualification	Qcat	6 categories
Age	AGE	continuous
Age squared	AGE^2	continuous
Length of time cont. employed	EMPMONcat	4 categories
Head of household	HOH	2 categories
Married	MARRIED	2 categories
No. of employees at workplace	SIZE	2 categories
Industry sector	INDcat	9 categories
Region	REGcat	12 categories
Gender	FEMALE	2 categories
Last pay same as usual	USGRS	2 categories
Interaction $\ln(X) \times$ occupation	LHE:SOCcat	-
Interaction $\ln(X) \times$ part-time	LHE:PT	-
Interaction $\ln(X) \times$ no.of employees	LHE:SIZE	-
Interaction $\ln(X) \times$ less than weekly	LHE:LTWK	-
Interaction $\ln(X) \times$ industry	LHE:INDcat	-

Table 2.1: Covariates employed in the imputation model.

2.2.5 Estimating the Distribution of Hourly Pay

The aim is to estimate the distribution of hourly pay $F_Y(y)$ and in particular the proportion of low paid employees in the population. Using imputation the estimator of interest $\hat{\theta}.$ is defined as

$$\hat{F}.(y) = \frac{\sum_{i \in s} w_i I(y_{.i} \leq y)}{\sum_{i \in s} w_i}, \tag{2.14}$$

where w_i denotes the survey weight for individual $i \in s$ for earnings data in the LFS. It is assumed in the following that the survey weights adequately compensate for selective nonresponse. In chapter 3 it is shown that the estimator in (2.14) is approximately unbiased for $F_Y(y)$ under the assumption of uniform within classes nonresponse. We may write the proportion in the population alternatively as

$$P = \frac{1}{N} \sum_{i \in U} z_i, \tag{2.15}$$

where z is the indicator variable, such that

$$z_i = I(y_i < t) = \begin{cases} 0 & \text{if } y_i \geq t \\ 1 & \text{if } y_i < t \end{cases} \quad \forall i \in U, \tag{2.16}$$

and t is a specified level of hourly pay such as the NMW. Based on the repeated imputations for estimating this proportion M binary variables, $z_{.m}$, are created, where m=1,..., M, according to each of the M imputations, indicating if person i in the sample earns less than the value t or not. We write

$$z_{.mi} = I(y_{.mi} < t) = \begin{cases} 0 & \text{if } y_{.mi} \geq t \\ 1 & \text{if } y_{.mi} < t \end{cases} \quad \forall m = 1,...,M, \tag{2.17}$$

or alternatively

$$z_{.mi} = \begin{cases} z_i = I(y_i < t) & \text{for } i \in r \\ \hat{z}_{mi} = I(\hat{y}_{mi} < t) & \text{for } i \in \bar{r} \end{cases} \quad \forall m = 1,...,M. \tag{2.18}$$

An estimator of the proportion of low paid employees based on the mth imputation, $m = 1,...,M$, is given by

$$\hat{P}_{\cdot m} = \frac{\sum_{i \in s} w_i z_{\cdot mi}}{\sum_{i \in s} w_i}. \tag{2.19}$$

Averaging these estimators gives the overall estimator of P as

$$\hat{P}_{\cdot} = \frac{1}{M} \sum_{m=1}^{M} \hat{P}_{\cdot m}. \tag{2.20}$$

In the following we present estimates of the proportion of employees being paid below the NMW rate. This particular proportion will be denoted $\hat{P}_1$.. Table 2.2 shows estimates of this proportion under hot deck imputation within classes for several quarters of the LFS. It can be seen that, because of the introduction of the NMW in April 1999, the estimated proportion of low paid employees for April to May 1999 was greatly reduced in comparison to March 1999, and continued to decrease over following quarters.

	Age: 18-21	**Age: 22+**
March 1999	3.1	5.7
April-May 1999	2.3	2.2
June-August 1999	2.0	1.6
Sept-Nov 1999	1.5	1.2
Dec 1999-Feb 2000	1.3	1.2
March-May 2000	1.6	0.6

Table 2.2: Estimated (weighted) percentages of employees earning below the NMW, $\hat{P}_1$., for several quarters of the LFS based on ONS hot deck imputation within classes (Stuttard and Jenkins, 2001).

To illustrate the effects of different estimation methods figure 2.5 presents estimates of the pay distribution based on several methods discussed above. The vertical line indicates the level of the NMW. Corresponding estimates of the proportion of low paid employees for the different methods are presented in table 2.3. The simplest approach is to use the derived variable making no use of values of the direct variable. It can be seen that this approach leads to the largest estimates of the proportion of low paid employees. The estimate for $\hat{P}_1$. is 6.6% under this method. The second largest estimate is obtained when the direct variable and the derived variable are combined, using the values of the direct value when observed and the values of the derived variable when the direct variable is missing. This gives an estimate of 4.1%. For regression imputation, imputing the predicted values without additional residuals as well as for hot deck imputation within classes, the estimate is greatly reduced to approximately 1.5% of low paid employees. We can see from the estimated distribution that regression imputation without residuals leads to underestimation in this part of the distribution (for further discussion see also section 6.3.5). Using only the observed values for the direct variable gives an estimate of 1.9%, which is assumed to be an overestimate since hourly paid employees tend to earn less than salaried employees. One reason to suppose that the results obtained from the imputation method are more accurate than those obtained from using the derived variable alone is that in figure 2.5 there is, as expected, a strong kink in the distribution at the NMW wage for the imputed direct variable, unlike for the derived variable.

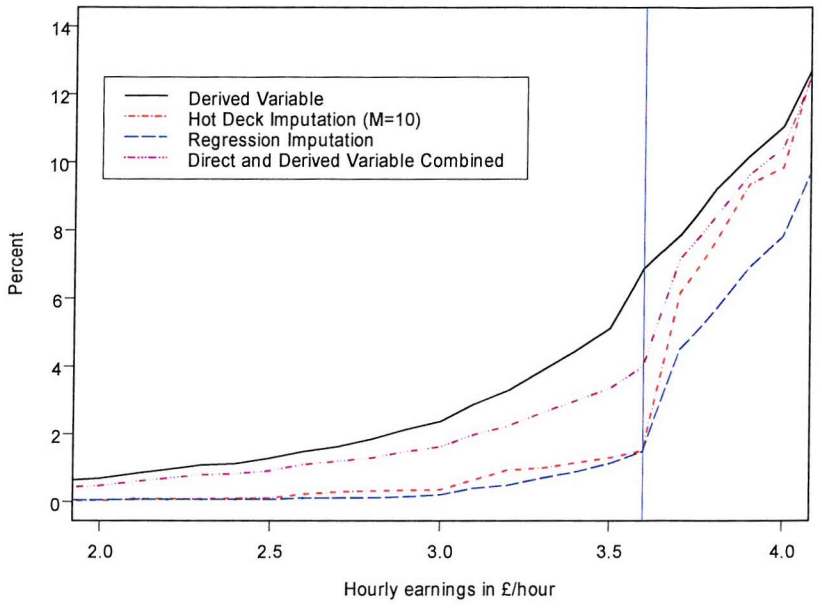

Figure 2.5: Estimated cumulative distributions of hourly earnings from £2 to £4 for 22+ age group for June-August 1999 (weighted), based on estimates using the derived variable only, hot deck imputation within classes with 10 imputations, regression imputation and a combination of direct and derived variable.

Method	$\hat{P}_1$.
Derived variable	6.6%
Direct and derived combined (direct variable where observed and derived variable otherwise)	4.1%
Hot deck imputation within classes, sampling donors without replacement (imputed direct variable)	1.6%
Regression imputation (direct variable where Y observed and predicted values otherwise)	1.5%
Observed values of direct variable only	1.9%

Table 2.3: Estimated (weighted) percentages of low paid employees for 22+ age group for June-August 1999.

Chapter 3

Evaluation of Hot Deck Imputation Within Classes

Different criteria can be used to evaluate imputation methods. The basic aim is to obtain a small or zero bias of the estimators of interest with an associated small inflation of variance. In addition, the performance of the imputation method under model misspecification and violations of underlying assumptions need to be analysed. Other aspects when evaluating an imputation method are the availability of valid variance estimation formulae and the implementation of the chosen procedure in practice.

This chapter investigates the performance of the hot deck imputation method, described in the previous chapter, both theoretically and empirically, focussing on the bias of the point estimator of interest, $\hat{P}$.. In section 3.1 the bias of the estimator, not taking into account LFS survey weights, is investigated theoretically. In section 3.2 we allow for fixed survey weights. It should be noted that the following theory does not take into account stratification, clustering into households or the complex poststratification procedure allowing for survey weights being dependent on the sample. Allowing for clustering into households would require taking account of the dependency of imputed values in the theoretical derivations. In section 3.3 a simulation study is carried out to evaluate the hot deck imputation method and the performance of the resulting point estimator

empirically. In section 3.4 the performance of the imputation method is evaluated under alternative assumptions, in particular under misspecification of the imputation model. The impact of the choice of imputation classes is also investigated. In addition, the effects of different specifications of the imputation method when applying the method to LFS data are analysed. Some concluding remarks are given in section 3.5. The variance of the point estimator and variance estimation will be discussed in chapter 4. In chapter 5 further investigations are carried out to evaluate the hot deck imputation method, in particular in comparison to other forms of predictive mean matching imputation, such as nearest neighbour imputation, and a weighting method. There, the impact of the different choices of the number of repeated imputations, M, as well as differences in stochastic and deterministic imputation methods are analysed.

3.1 Evaluation of Unweighted Point Estimator

3.1.1 Theoretical Framework and Definitions

In this section the properties of the estimator $\hat{P}.$ are analysed theoretically not taking into account LFS survey weights. It is shown that $\hat{P}.$ is an approximately unbiased estimator of P under hot deck imputation within classes assuming ignorable nonresponse, using notation from section 2.2.5. To facilitate our discussion the following notation is introduced.

Definition 3.1

r = set of respondents on Y, the variable of interest subject to nonresponse, and $r = 1,\ldots, n_r$, where n_r is the number of respondents on Y in s.

$\bar{r}$ = set of non-respondents on Y, $\bar{r} = 1,\ldots, n_{\bar{r}}$, where $n_{\bar{r}}$ is the number of nonrespondents, such that $n = n_r + n_{\bar{r}}$

B_{Uk} = imputation class based on the population U, where $k = 1,\ldots, K$

N_k = number of employees in imputation class k in population, such that $\sum_k N_k = N$

π_k = probability of response in class B_{Uk}

$\hat{\pi}_k$ = n_{rk} / n_k in sample s, estimator of π_k

B_k = imputation class based on the sample s, where k=1,... , K

n_k = number of employees in imputation class k in s, such that $\sum_k n_k = n$

n_{rk} = number of respondents on variable Y in imputation class k

$B_{k(i)}$ = imputation class which contains person i

$n_{rk(i)}$ = number of respondents on variable Y in imputation class $B_{k(i)}$

M = number of imputations (e.g. M=10)

The estimator $\hat{P}.$, not taking into account the survey weights, is of the form

$$\hat{P}. = \frac{1}{M}\sum_{m=1}^{M}\hat{P}_{.m}, \text{ where } \hat{P}_{.m} = \frac{1}{n}\sum_{i\in s} z_{.mi}, \qquad (3.1)$$

$m = 1,...,M$ and n denotes the number of employees in the sample. To investigate the sampling distribution of the point estimator $\hat{P}.$ we take into account the sampling mechanism, denoted D, the response mechanism, R, and the imputation variability, I. We do not refer to the distribution ξ, where ξ defines an assumed imputation model. This additional distribution will be discussed in chapter 4. The sampling mechanism describes the way the sample s is drawn from the population U. For simplicity, simple random sampling (SRS) will be assumed. The actual complex design was described in section 2.1.2. This design involves equal probabilities of selection and so we do not expect this simplifying assumption to affect the bias of the estimator. We may expect some impact of the complex design on the variance of the estimator (Holmes and Skinner, 2000) but this effect is not our main concern here and is thus ignored for simplicity. The response mechanism expresses the unknown conditional probability that the subset r, the set of respondents, responds. In the following an ignorable nonresponse mechanism is assumed making the assumption of uniform nonresponse within classes. Since the imputation classes are formed using a regression model, where Y is regressed on the covariates X and W, the assumption for the nonresponse used here indirectly takes into account that the probability of response does not depend on Y given the auxiliary information X and W. Let π_k be the probability of response in each imputation class, for $k = 1,...,K$, assumed equal for all individuals in B_k. A formal problem

is that B_k is sample dependent and it is implausible that the response probability of an individual should be sample-dependent in this way. However, for large samples B_k is approximately equal to $B_{Uk} \cap s$ under the assumption of ignorable nonresponse, where B_{Uk} is the corresponding set in the population. The classes B_k are defined with respect to the predicted values based on a vector $\hat{\beta}$, i.e. based on $\hat{y}_{regi} = \exp(\eta_i \hat{\beta})$ in s. In the same way the classes B_{Uk} are defined with respect to the predicted values based on a vector β, i.e. based on $y_{regi} = \exp(\eta_i \beta)$ in U, where for large n the estimator $\hat{\beta}$ converges to a vector β. We thus assume that the response probability π_k is constant within B_{Uk} and use $B_k = B_{Uk} \cap s$ as an approximation. We also need to take into account the imputation variability (I) since the imputation method uses a random component and not a predicted value for imputing the missing observations. For the unweighted case the donors are selected within each class by simple random sampling (SRS) with replacement. With replacement has the advantage that the imputed values are independent for all $j \in \bar{r}$ and $m = 1,...,M$. The expressions E_D, E_R and E_I denote the expectation operators with respect to D, R and I. To facilitate our discussion we need to introduce some notation, which will be used in the following sections.

Definition 3.2

Within each imputation class let

$$\bar{z}_{Uk} = \frac{1}{N_k} \sum_{i \in B_{Uk}} z_i \tag{3.2}$$

be the proportion of low paid employees in imputation class B_{Uk} in the population,

$$\bar{z}_{sk} = \frac{1}{n_k} \sum_{i \in B_k} z_i \tag{3.3}$$

be the corresponding proportion in imputation class B_k in the sample under full response,

$$\bar{z}_{rk} = \frac{1}{n_{rk}} \sum_{i \in B_k \cap r} z_i \tag{3.4}$$

be the proportion of low paid employees in the sample amongst the respondents and let

$$\bar{z}_{\cdot sk} = \frac{1}{M}\sum_{m=1}^{M} \bar{z}_{\cdot skm}, \text{ where} \tag{3.5}$$

$$\bar{z}_{\cdot skm} = \frac{1}{n_k}\sum_{i\in B_k \cap s} z_{\cdot mi} = \frac{1}{n_k}\left[\sum_{i\in B_k \cap r} z_i + \sum_{i\in B_k \cap \bar{r}} \hat{z}_{mi}\right] \quad \forall m=1,\ldots,M, \tag{3.6}$$

be the proportion of low paid employees in imputation class B_k in the sample subject to repeated imputation. When considering the expectation according to imputation we use the following definition

$$\bar{z}_{nk(i)} = \frac{1}{n_{nk(i)}}\sum_{j\in B_{k(i)}\cap r} z_j = \frac{\sum_{j\in B_{k(i)}} z_j I(j\in r)}{\sum_{j\in B_{k(i)}} I(j\in r)}, \tag{3.7}$$

which is the proportion of respondents with z_j=1, with unit j being in the same class as unit i.

3.1.2 Unbiasedness of Point Estimator

In this section it is shown that the point estimator $\hat{P}_{\cdot}$ is an approximately unbiased estimator for P with respect to D, R and I under the assumptions outlined in section 3.1.1. In particular, in large samples the bias of such an estimator is shown to be negligible. Result 3.1 and its proof are relevant for deriving valid variance formulae in section 4.2.

Result 3.1

Under the assumptions of (i) equal probability sampling, (ii) uniform nonresponse within classes and (iii) n_{nk}, the number of respondents in class B_k, is large for all k,

the unweighted estimator $\hat{P}.$, obtained under the imputation method described in section 2.2.3 selecting donors by simple random sampling with replacement, is approximately unbiased for P, i.e.

$$E_{DRI}(\hat{P}.) \doteq P \tag{3.8}$$

Proof

$$E_{DRI}(\hat{P}.) = E_{DRI}\left[\frac{1}{M}\sum_{m=1}^{M}\hat{P}._m\right] = \frac{1}{M}\sum_{m=1}^{M}E_{DRI}(\hat{P}._m) \doteq P, \text{ if } E_{DRI}(\hat{P}._m) \doteq P \ \forall m$$

To show this:

$$E_{DRI}(\hat{P}._m) = E_{DRI}\left[\frac{1}{n}\left[\sum_{i\in r} z_i + \sum_{i\in \bar{r}} \hat{z}_m\right]\right] = E_{DR}\left[\frac{1}{n}\left[\sum_{i\in r} z_i + \sum_{i\in \bar{r}} E_I(\hat{z}_m)\right]\right]$$

where $E_I(\hat{z}_m) = pr(\hat{z}_m = 1) = pr(\hat{y}_m < t))$

= probability of choosing person j in the same imputation class as i with $y_j < t$

= proportion of respondents with $y_j < t$, i.e. $I(y_j < t) = 1$

= proportion of respondents with $z_j = 1$

$$= \frac{1}{n_{r k(i)}}\sum_{j\in B_{k(i)}\cap r} z_j = \frac{\sum_{j\in B_{k(i)}} z_j I(j\in r)}{\sum_{j\in B_{k(i)}} I(j\in r)} = \bar{z}_{r k(i)} \quad \text{for all } m=1, \ldots, M, \text{ using (3.7)}$$

where $B_{k(i)}$ denotes imputation class B_k, which contains person i, $i \in \bar{r}$, and $n_{r k(i)}$ denotes the number of respondents on Y in imputation class $B_{k(i)}$.

It follows that

$$E_{DRI}(\hat{P}._m) = E_{DR}\left[\frac{1}{n}\sum_{i\in s}\left\{I(i\in r)z_i + I(i\in \bar{r})\bar{z}_{r k(i)}\right\}\right] = E_{DR}\left[\frac{1}{n}\sum_{i\in s}\left\{I(i\in r)(z_i - \bar{z}_{r k(i)}) + \bar{z}_{r k(i)}\right\}\right]$$

$$= E_{DR}\left[\frac{1}{n}\sum_{i\in s}\bar{z}_{r k(i)}\right] \qquad \text{since} \quad \frac{1}{n}\sum_{i\in s} I(i\in r)(z_i - \bar{z}_{r k(i)}) = \frac{1}{n}\sum_{i\in r}(z_i - \bar{z}_{r k(i)})$$

$$= \frac{1}{n}\sum_k \sum_{i\in B_k \cap r} (z_i - \bar{z}_{rk}) = \frac{1}{n}\sum_k \left(\sum_{i\in B_k \cap r} z_i - n_{rk}\bar{z}_{rk} \right) = 0$$

$$= E_{DR}\left[\frac{1}{n}\sum_k \sum_{i\in B_k} \bar{z}_{rk}\right] = E_D\left[\frac{1}{n}\sum_k n_k E_R(\bar{z}_{rk})\right]$$

where $E_R(\bar{z}_{rk}) = E_R\left[\frac{1}{n_{rk}} \sum_{i\in B_k \cap r} z_i\right] = E_R\left[\frac{\sum_{i\in B_k} z_i I(i\in r)}{\sum_{i\in B_k} I(i\in r)}\right] \doteq \frac{\sum_{i\in B_k} z_i E_R(I(i\in r))}{\sum_{i\in B_k} E_R(I(i\in r))}$

assuming n_{rk} large for all k

Note: we can write $E(x/y) \doteq E(x)/E(y)$ if var(x) and var(y) small (Cochran, 1977, p. 31, theorem 2.5; see also the Taylor approximation in section 3.2, equation (3.14)).

$$= \frac{\sum_{i\in B_k} \pi_k z_i}{\sum_{i\in B_k} \pi_k} = \frac{1}{n_k}\sum_{i\in B_k} z_i = \bar{z}_{sk} \qquad (3.9)$$

$$= E_D\left[\frac{1}{n}\sum_k n_k \bar{z}_{sk}\right] = E_D\left[\frac{1}{n}\sum_k \sum_{i\in B_{Uk}} z_i I(i\in s)\right] = E_D\left[\frac{1}{n}\sum_{i\in U} z_i I(i\in s)\right] = \frac{1}{N}\sum_{i\in U} z_i = \bar{z}_U = P$$

❑

Using Taylor linearisation as in (3.13) the first order Taylor approximation used in the above proof is of order $1/n$ and it follows that $E_{DRI}(\hat{P}.) = P + O(n^{-1})$. The approach used here to prove that the imputed estimator is approximately unbiased is a design-based approach. Similar approaches have been used by Rao and Shao (1992) and Rao (1996), who show that under uniform and uniform within classes nonresponse using hot deck imputation the imputed estimator for the population total is approximately unbiased. Alternatively, a model-based approach can be used specifying a superpopulation model ξ, which is also referred to as an imputation model (Rao, 1996). The model-based approach as well as differences between the design-based and model-based approach in the presence of imputed data with particular reference to variance estimation is discussed in chapter 4. Under the model-based approach the imputation model is stated explicitly.

It is possible to make a weaker assumption about the nonresponse, i.e. MAR nonresponse (Lee, Rancourt and Särndal, 2000; Deville and Särndal, 1994). The cases of regression and ratio imputation are discussed by Rao (1996). He shows that if the respective imputation model holds and assuming MAR nonresponse the resulting imputed estimator for the population mean is approximately design-model (or $D\xi$) unbiased. The case of regression imputation is also discussed by Rao (2001) and Sitter and Rao (1997). Under uniform nonresponse or uniform nonresponse within classes the resulting estimators are approximately design-unbiased. Under hot deck imputation Rao (1996) shows that the imputed estimator is approximately design-model unbiased under an appropriate superpopulation model assuming MAR. We conclude that in general we either need the more restrictive assumption of uniform response or uniform within classes nonresponse without the imputation model or the condition of MAR with the model (Rao, 1996; Deville and Särndal, 1994; Rao, 2001).

3.2 Evaluation of Point Estimator Taking Account of Fixed Survey Weights

3.2.1 Theoretical Framework and Definitions

In this section the survey weights w_i are taken into account, such that the point estimator is of the form

$$\hat{P}_{\cdot} = \frac{1}{M}\sum_{m=1}^{M}\frac{\sum_{i\in s} w_i z_{mi}}{\sum_{i\in s} w_i}, \tag{3.10}$$

as described in section 2.2.5. It is assumed that the survey weights adequately compensate for unit nonresponse in the sample s in the following sense. The subscript D, which previously referred to the sampling distribution, here refers to the sampling and the (unit) nonresponse distribution. We denote $v_i = E_D(I(i \in s))$, such that v_i is the multiplication of the inclusion probabilities and the probability of (unit) response for unit i. The survey weights are assumed to be the reciprocal of the probabilities of being in sample s, i.e. $w_i = 1/v_i$. Thus, it is assumed that the weights compensate for differential nonresponse via a standard kind of Horvitz-Thompson weighting. The survey weights are regarded as fixed in the population, such that we can write

$E_D(\sum_{i\in s} w_i z_{mi}) = \sum_{i\in U} w_i z_{mi} E_D(I(i\in s))$. This assumption is an approximation since the weights are determined through the raking adjustment and are therefore strictly speaking sample dependent. This approximation may be justified since the weights are centred around a central point and the variance of the weights may be regarded as small. The mean and the median of the weights are approximately 1450 for the March-May 2000 quarter. The interquartile range is between 1300 and 1600 and 90% of all weights are within the range of approximately 1100 and 1800 which indicates the concentration around the central point. Regarding the survey weights as fixed is also discussed in Holmes and Skinner (2000, p. 11). They suggest a way of allowing for the poststratification adjustment. There, the weighted estimator is approximated by a poststratified estimator and the usual variance estimator for a post-stratified estimator is used to obtain a variance estimator.

For the weighted case the selection of donors in the imputation procedure depends on the survey weights in the following way as described in section 2.2.3. Respondent i, $i \in B_k \cap r$, is selected as a donor for nonrespondent j, $j \in B_k \cap \bar{r}$, with probability $w_i / \sum_{i\in B_k \cap r} w_i$. The selection is carried out with replacement. When considering the expectation according to imputation we therefore use the following definition

$$\bar{z}_{rk(j)} = \frac{\sum_{i\in B_{k(j)}} w_i z_i I(i\in r)}{\sum_{i\in B_{k(j)}} w_i I(i\in r)}, \qquad (3.11)$$

which is the weighted proportion of respondents with z_i =1 and unit i being in the same class as unit j. Rao and Shao (1992) and Rao (1996) also suggest the sampling of donors with probabilities depending on the survey weights under hot deck imputation and show that the resulting imputed estimator for the population total is approximately unbiased under uniform and uniform within classes nonresponse. To facilitate our discussion we introduce the following notation.

Definition 3.3

Similar to the unweighted case we define the following terms for the weighted case. Let

$$\bar{z}_{Uk} = \frac{1}{N_k}\sum_{i\in B_{Uk}} z_i, \quad \bar{z}_{sk} = \frac{\sum_{i\in B_k} w_i z_i}{\sum_{i\in B_k} w_i}, \quad \bar{z}_{rk} = \frac{\sum_{i\in B_k} w_i z_i I(i\in r)}{\sum_{i\in B_k} w_i I(i\in r)} \text{ and}$$

$$\bar{z}_{\cdot sk} = \frac{1}{M}\sum_{m=1}^{M} \bar{z}_{\cdot skm} \text{, where } \bar{z}_{\cdot skm} = \frac{\sum_{i \in B_k} w_i z_{\cdot mi}}{\sum_{i \in B_k} w_i} \quad \forall\ m = 1,\ldots,M, \tag{3.12}$$

be the proportions of low paid employees in the kth class in the population, in the sample, among the respondents and in the sample including imputed values.

Remarks

Before carrying out the analysis it is useful to make some general remarks, which are also used for deriving the variance in chapter 4.

1) To approximate expectations such as $E(\bar{x}_s\ \bar{y}_s)$ and $E(\bar{x}_s / \bar{y}_s)$, where $\bar{x}_s$ denotes the estimate of the mean of a random variable X based on sample s and $\bar{y}_s$ the quantity for a random variable Y, we use the first order Taylor series expansion of $\bar{x}_s\ \bar{y}_s$ (or $\bar{x}_s / \bar{y}_s$):

$$g(\bar{x}_s, \bar{y}_s) = \bar{x}_s\ \bar{y}_s \doteq \bar{x}_U\ \bar{y}_U + (\bar{x}_s - \bar{x}_U)\frac{\partial g}{\partial \bar{x}_s}\Big|_{\bar{x}_s = \bar{x}_U, \bar{y}_s = \bar{y}_U} + (\bar{y}_s - \bar{y}_U)\frac{\partial g}{\partial \bar{y}_s}\Big|_{\bar{x}_s = \bar{x}_U, \bar{y}_s = \bar{y}_U}, \tag{3.13}$$

The terms $(\bar{x}_s - \bar{x}_U)$ and $(\bar{y}_s - \bar{y}_U)$ are of order $O_p(n^{-1/2})$, whereas the remainder $(\bar{x}_s - \bar{x}_U)(\bar{y}_s - \bar{y}_U)$ is of smaller order, i.e. of order $O_p(n^{-1})$. The order of approximation used here is therefore $O_p(n^{-1})$.

We apply the expectation to the approximation in (3.13):

$$E(\bar{x}_s\ \bar{y}_s) \doteq E(\bar{x}_U\ \bar{y}_U) = \bar{x}_U\ \bar{y}_U = E(\bar{x}_s)E(\bar{y}_s), \tag{3.14}$$

where U refers to the quantity in the population. In the context used here, the Taylor series approximation requires the number of respondents and the number of units in class B_k to be reasonably large for all k. This requirement is realistic for the application to the LFS considered here.

Example: $E_D(\sum_{i \in B_k} w_i z_i \bar{z}_{sk}) = E_D(\bar{z}_{sk} \sum_{i \in B_k} w_i z_i) \doteq E_D(\bar{z}_{sk}) E_D(\sum_{i \in B_k} w_i z_i) = \bar{z}_{Uk} \sum_{i \in B_{Uk}} z_i$ (3.15)

3.2.2 Unbiasedness of Point Estimator Taking Account of Fixed Survey Weights

In the following we show that the estimator $\hat{P}.$ is approximately unbiased taking into account the weighting scheme of the LFS regarding the survey weights as fixed.

Result 3.2

Under the assumptions of (i) equal probability sampling, (ii) uniform nonresponse within classes and (iii) the number of respondents and the number of units in class B_k are large for all k, the weighted estimator $\hat{P}.$, obtained under the imputation method described in section 2.2.3 selecting donors with replacement, is approximately unbiased for P, i.e.

$$E_{DRI}(\hat{P}.) \doteq P \tag{3.16}$$

Proof

It is enough to show that $E_{DRI}(\hat{P}_{.m}) \doteq P \ \ \forall m$

$$E_{DRI}(\hat{P}_{.m}) = E_{DRI}\left((\sum_{i\in s} w_i)^{-1}\sum_{i\in s} w_i\, z_{.mi}\right)$$

$$\doteq \frac{1}{N} E_{DR}\left(\sum_k \sum_{i\in B_k}\{I(i\in r)w_i z_i + (1-I(i\in r))w_i \bar{z}_{rk}\}\right) = \frac{1}{N} E_{DR}(\sum_k \bar{z}_{rk} \sum_{i\in B_k} w_i)$$

Note: $E_I(\hat{z}_{mi}) = \bar{z}_{rk(i)}$ using definition (3.11) and

$$\bar{z}_{rk}\sum_{i\in B_k} I(i\in r)w_i = \sum_{i\in B_k} I(i\in r)w_i z_i$$

$$\doteq \frac{1}{N} E_D\left(\sum_k (\sum_{i\in B_k} w_i)(E_R(\sum_{i\in B_k} w_i I(i\in r)))^{-1} E_R(\sum_{i\in B_k} w_i z_i I(i\in r))\right)$$

$$= \frac{1}{N} E_D\left(\sum_k (\sum_{i\in B_k} w_i)(\sum_{i\in B_k} w_i\pi_k)^{-1}\sum_{i\in B_k} w_i z_i \pi_k\right) = \frac{1}{N} E_D(\sum_k \sum_{i\in B_k} w_i z_i) = P$$

❑

3.3 Simulation Study Evaluating the Point Estimator

The aim of this simulation study is to assess the performance of the estimator $\hat{P}.$ obtained under the proposed hot deck imputation method within classes empirically. The main emphasis is on the bias of the point estimator. In addition, the robustness of the hot deck method under model misspecification and under different specifications of the imputation classes is investigated. The variance of the estimator $\hat{P}.$ as well as corresponding variance estimators are discussed in chapter 4.

For the simulation study A =1000 independent samples $s^{(a)}$, $a=1,...,A$, are generated containing values for Y, X, W and I using data from the LFS. A sampling procedure is simulated by bootstrapping an original dataset from the LFS. Nonresponse is introduced according to three different nonresponse mechanisms, uniform, uniform within classes and MAR nonresponse. The performances of two imputed point estimators $\hat{P}_1.$, the proportion of employees earning below the NMW (t=£3.60), and $\hat{P}_2.$, the proportion of employees earning between £3.60 and £5, are evaluated. For simplicity the survey weights of the LFS are not taken into account in the simulation study.

3.3.1 Generating the Population and Samples

The aim is to create A samples each of size n, containing the response variable Y, X and other relevant auxiliary variables W. Strictly speaking it would be necessary to generate a population U, including information on these variables for all $i \in U$ and draw h samples from this population to reflect the sampling variability. However, this population would need to be very large (e.g. N=20 million), such that it was decided to reflect sampling variability by selecting units from the original LFS sample with replacement, that is to adopt a Bootstrap approach, and to create samples $s^{(a)}$, of the same size as the original sample with n=15,000. We therefore assume that an infinite population exists from which samples are drawn independently at each replication, which seems reasonable because of the small sampling fraction of the LFS. The March-May 2000 quarter was chosen as the original LFS sample. All employees younger than 18 and employees that belong to certain industry groups are not taken into account since the NMW legislation does not apply to

these groups. The very small number of cases with missing values on X or W were omitted. An alternative approach for generating the data would be to regard the original sample as the population and to sample in each replication from this population. To create this population it would be possible to keep all the observed values in this original sample and to generate only the missing values for the variable Y. However, it would be necessary to use a much smaller sample size, e.g. n=1000, which might lead to unrealistic results. This approach was therefore not used.

By selecting a bootstrap sample $s^{(a)}$ the values for the covariates $W^{(a)}$ for all units i =1,... , n are obtained directly from the LFS sample. The values for $\ln(Y)$, $\ln(X)$ and I are generated from models as follows. The variable $\ln(Y)$ needs to be generated, since $\ln(Y)$ has missing values in the original sample. The values for $\ln(Y)$ are generated for each unit according to a model taking into account the values for $\ln(X)$ and W and adding on random residuals $\hat{\varepsilon}^{(a)}$, i.e. we have

$$\ln(y_i^{(a)}) = \eta_i^{(a)}\hat{\beta}_r + \hat{\varepsilon}_i^{(a)} \quad , i=1,\dots, n, \tag{3.17}$$

where the coefficient vector $\hat{\beta}_r$ is estimated from the original LFS sample based on respondents only and is held fixed in the simulation. The model uses all the variables specified in section 2.2.4. The number of respondents in the original sample is approximately 7000. The distribution of the residuals in the original sample was found to be approximately normal with mean zero and variance $\hat{\sigma}_\varepsilon^2$ =0.038. The residuals in (3.17) are chosen to be random draws from a normal distribution with $\hat{\varepsilon}_i^{(a)} \sim N(0,\hat{\sigma}_\varepsilon^2)$.

The variable $\ln(X)$ is generated to avoid duplications of units reporting the same values on $\ln(X)$ and W, which could be regarded as an unrealistic condition since the imputation method relies on the predicted values of $\ln(Y)$ based on $\ln(X)$ and W. The model generating $\ln(X)$ is of the form

$$\ln(x_i^{(a)}) = \mu_i^{(a)}\hat{\lambda}_r + \hat{\varepsilon}_{xi}^{(a)}, \; i=1, \dots, n, \tag{3.18}$$

where $\mu_i^{(a)}$ is a row-vector of functions of the covariates W, $\hat{\lambda}_r$ is a vector of coefficients and the residuals $\hat{\varepsilon}_{xi}^{(a)}$ are random draws from a normal distribution. The parameters necessary for this model are estimated based on respondents in the original LFS dataset. The variables included in

this model are given in table 3.1. Sample $s^{(a)}$ is now a complete dataset, containing $\ln(Y)^{(a)}$, $\ln(X)^{(a)}$ and $W^{(a)}$ for all i=1, ... , n.

Name of Variable	Abbreviation	Continuous/ Categorical
Major occupation group	SOC	9 categories
Age	AGE	continuous
Age squared	AGE^2	continuous
Highest qualification	Q	6 categories
Part-time	PT	2 categories
Length of time continuously employed	EMPMON	4 categories
Paid less than weekly	LTWK	2 categories
Head of household	HOH	2 categories
Married	MARRIED	2 categories
Number of employees at workplace	SIZE	2 categories
Gender	FEMALE	2 categories
Interaction occupation × part-time	SOC:PT	-

Table 3.1: Variables included in the linear regression model generating $\ln(X)$ in the simulation study.

3.3.2 Simulating Nonresponse

The values of the response variable I are generated based on three nonresponse mechanisms, uniform, uniform within classes and MAR nonresponse. A nonignorable nonresponse mechanism is considered in chapter 6.

i.) Uniform Nonresponse Mechanism

Uniform nonresponse means that the response probability is constant for all units $i \in s^{(a)}$ assuming independent response across sample units, i.e. $pr(i \in r^{(a)}) = c$ for all $i \in s^{(a)}$, where

$0 < c < 1$. In the simulation study the constant c is set to 0.43, since the response rate in the original dataset is approximately 43%. Bernoulli sampling is used to create a uniform response mechanism. We take $\delta_1, ..., \delta_n$ independent draws of a Unif(0,1) random variable in each sample $s^{(a)}$, where Unif(0,1) defines the random variable with a uniform distribution on the interval (0,1). If a person belongs to the set of respondents r or to the nonrespondents $\bar{r}$ is decided by

$$\text{if } \delta_i < c \Rightarrow i \in r \text{, otherwise } i \in \bar{r}, \tag{3.19}$$

such that $pr(i \in r^{(a)}) = pr(\delta_i < c) = c$ for all $i \in s^{(a)}$ and the events "$i \in r^{(a)}$" and "$j \in r^{(a)}$" are independent for $i \neq j$. The number of respondents is a binomially distributed random variable with parameters n and c.

ii.) Uniform within Imputation Classes

To introduce nonresponse based on the assumption used in the theoretical derivation, i.e. uniform response within classes, seven classes are constructed according to the range of the predicted values of $Y^{(a)}$, which are obtained using the auxiliary variables in $s^{(a)}$ and the coefficients based on the respondents in the original sample, i.e. $\hat{\beta}_r$. The boundaries of the classes are held fixed. They reflect £1.5 pay bands. The probability of response for each class is fixed, such that the probabilities are approximately the same as the probabilities in the original sample (table 3.2). That means that in response classes reflecting higher earnings, e.g. in class 6 and 7, the probability of response is small, whereas in response classes containing mostly low earners the response rate is higher. This is the case since employees with an hourly rate are more likely to be low paid. The overall response rate is approximately 43%. To select units for nonresponse Bernoulli sampling is performed in each class.

Response class	1	2	3	4	5	6	7
prob. of response	0.78	0.75	0.56	0.41	0.31	0.20	0.08

Table 3.2: Fixed probabilities of response in response classes, based on estimated average response probabilities in each class in the original LFS sample.

When applying imputation, the imputation classes are defined by using the predicted values from the regression carried out in sample $s^{(a)}$ (see formula (3.23)), where the values of the auxiliary variables and the coefficients depend on the values in sample $s^{(a)}$. The boundaries of the imputation classes are defined in the same way as the boundaries for the response classes. Note that the response classes and the imputation classes can differ slightly, in the sense that if a unit i belongs to response class k it may not necessarily belong to imputation class k and vice versa. This difference in response and imputation classes also enables us to test robustness against the assumption that response rates are strictly constant in each imputation class.

iii.) Nonresponse Following the MAR Assumption

To implement a nonresponse mechanism that follows the MAR assumption the values for I in $s^{(a)}$ are generated using a model such that the probability of response depends on the auxiliary variables X and W and therefore varies for each individual. Several logistic regression models were fitted in the original sample of the form

$$\log\left(\frac{\hat{p}_i}{1-\hat{p}_i}\right) = \eta_i \hat{\varphi}_r \quad \text{for all } i = 1, \ldots, n, \tag{3.20}$$

where η is a vector of functions of the derived variable and other covariates W. Note that the auxiliary variables W used in the model (3.20) are not necessarily the same as in the regression model (3.17) since it also includes variables that are considered to be predictors of the nonresponse. More information on the variables included in the logistic regression model is given in section 3.3.4. The estimated probability of response for unit i in sample $s^{(a)}$ is obtained using information on $X^{(a)}$ and $W^{(a)}$ and the coefficients $\hat{\varphi}_r$ obtained from the original sample, such that

$$\hat{p}_i^{(a)} = \frac{\exp(\eta_i^{(a)} \hat{\varphi}_r)}{1+\exp(\eta_i^{(a)} \hat{\varphi}_r)}. \tag{3.21}$$

In each sample $s^{(a)}$ we generate $\delta_1, \ldots, \delta_n$ independent realisations of a Unif(0,1) random variable. Response and nonresponse is decided by:

$$\text{if } \delta_i^{(a)} < \hat{p}_i^{(a)} \Rightarrow i \in r^{(a)}, \text{ otherwise } i \in \bar{r}^{(a)}, \text{ for } i = 1, \ldots, n. \tag{3.22}$$

The overall response rate in each sample $s^{(a)}$ was found to be approximately 43%. Carrying out the simulation study it was found that the average probability of response in each imputation class is approximately as in table 3.3. We can see that on average the probabilities of response under the logistic model are very similar to the estimates obtained in the original sample. We therefore conclude that the logistic model on average reflects the probabilities of response for each class and the structure of nonresponse well, since the higher the hourly earnings the lower is the response rate. (The results reported here refer to the logistic model A3 as described in section 3.3.4.)

Imputation class	1	2	3	4	5	6	7
prob. of response	0.82	0.71	0.55	0.41	0.32	0.18	0.08

Table 3.3: Average probability of response in each imputation class obtained using a logistic regression model.

3.3.3 Imputation

For imputing the missing values in Y using a form of predictive mean matching a regression model of the form

$$\ln(y_i^{(a)}) = \eta_i^{(a)} \hat{\beta}^{(a)} + \hat{\varepsilon}_i^{(a)}, \ \forall i \in r^{(a)}, \tag{3.23}$$

is fitted in sample $s^{(a)}$ based on respondents only. Initially the same covariates are included as in model (3.17). Units are allocated to imputation classes based on the predicted values in sample $s^{(a)}$,

$$\hat{y}_{regi}^{(a)} = \exp(\eta_i^{(a)} \hat{\beta}^{(a)}), \ \forall i \in s^{(a)}. \tag{3.24}$$

Hot deck imputation within classes is carried out multiple times with M=10, selecting donors with and without replacement. Both imputed point estimators $\hat{P}_1.$ and $\hat{P}_2.$ are calculated and steps 3.3.1-3.3.3 are repeated A =1000 times.

3.3.4 Modelling the Nonresponse in the LFS

Several logistic regression models are fitted in the original LFS sample to model the nonresponse and to determine the factors that are predictors of the nonresponse. The estimated coefficients are necessary to define the models generating nonresponse following MAR. In the original sample 14 variables, listed in table 3.4 are found to be significant in determining the nonresponse on the variable Y. These are variables that predict hourly pay as well as variables that are related to the response process, such as derived hourly pay, occupation, part-time and full-time, proxy response, industry section, highest qualification, region etc. It should be noted that derived hourly pay is the most significant variable followed by occupation. The model, using the derived variable as a continuous variable and including variables as given in table 3.4, is referred to as model A1. Testing the significance of squared terms it is found that the squared term for the derived hourly pay variable is highly significant such that this term is included in the model, denoted model A2. In addition, two significant two-way interaction terms with the derived variable are found and included in the model, denoted model A3. The table of coefficients for model A3 is given in the appendix A3.1.

The derived variable is also categorised into 11 categories depending on the range of values. A logistic regression model is fitted including the derived variable as a categorical variable and in addition all other variables given in table 3.4. As before the derived hourly pay variable is the most significant variable, followed by occupation. This model is referred to as model B. A third model is fitted including all variables as in model A1 apart from the variable derived hourly pay, despite the fact it is the most significant variable. This model is called model C.

Name of Variable	Abbreviation by ONS	Continuous/ Categorical
Ln of derived hourly pay= $\ln(X)$	LHE	continuous
Major occupation group	SOCcat	9 categories
Part-time	PT	2 categories
Paid less than weekly	LTWK	2 categories
Length of time cont. employed	EMPMONcat	4 categories
Additions to basic pay	ADDTBP	2 categories
Proxy response	PROXY	3 categories
Head of household	HOH	2 categories
Last pay same as usual	USGRS	2 categories
Sex	FEMALE	2 categories
No. of employees at workplace	SIZE	2 categories
Highest qualification	Qcat	6 categories
Industry sector	INDcat	9 categories
Region	REGcat	12 categories

Table 3.4: Variables employed in logistic regression model A1.

3.3.5 Performance and Evaluation

To evaluate the performances of the two point estimators the values for P_1, the proportion of low paid employees in the population, and P_2, the proportion of employees earning between the NMW and £5 in the population, are necessary. The values for P_1 and P_2 are derived using the average of $P_1^{(a)}$ and $P_2^{(a)}$ obtained based on the complete variable $Y^{(a)}$ over the A =1000 samples, i.e.

$$P_g = \sum_{a=1}^{A} P_g^{(a)}, \tag{3.25}$$

where g=1, 2. The values found for P_1 and P_2, using a fixed number generator for the iterations, were found to be P_1 =0.056 and P_2 =0.185. Alternatively, it is possible to generate a population by creating a very large Bootstrap sample, generating the complete variable Y in this population, and to obtain the values for P_1 and P_2 directly. This possibility was also explored and it was found that the results for the point estimators under this method are very similar to those using (3.25). To

assess the performances of both point estimators the bias and the relative bias are calculated. The estimators are denoted $\hat{P}_1.$ and $\hat{P}_2.$ respectively.

i.) Estimated Bias of the Point Estimator

$$\hat{Bias}(\hat{P}.) = E_s(\hat{P}.) - P = \frac{1}{A}\sum_{a=1}^{A}\hat{P}.^{(a)} - P, \tag{3.26}$$

where E_s is the average over all replications of the simulation.

ii.) Estimated Relative Bias of the Point Estimator

$$\hat{RB}(\hat{P}.) = 100 * \frac{E_S(\hat{P}.) - P}{P}. \tag{3.27}$$

iii.) Standard Error of the Bias Estimator

$$\hat{ste}(\hat{Bias}(\hat{P}.)) = \sqrt{V} / \sqrt{A}, \tag{3.28}$$

where $V = \frac{1}{A-1}\sum_{a=1}^{A}(\hat{P}.^{(a)} - \bar{P})^2$ and $\bar{P} = \frac{1}{A}\sum_{a=1}^{A}\hat{P}.^{(a)}$.

3.3.6 Simulation Results for the Point Estimator

Table 3.5 and 3.6 present the results for (estimated) bias and relative bias of the two point estimators under hot deck imputation within classes where donors are selected with and without replacement respectively. The biases of both point estimators based on M=10 imputed values are small with under 2% for almost all cases. Only under uniform nonresponse is the bias for both point estimators non-significant for selecting donors with and without replacement. In all other cases the bias is significant, however, it is very small. Under MAR nonresponse the bias varies between 0.5% and 2% depending on the model chosen to generate the nonresponse. We conclude that there is an indication of bias using the hot deck imputation method within classes. This might be related to the choice and width of the imputation classes. However, the bias seems to be very small. The performance of the point estimator $\hat{P}.$ when imputation without replacement is used is very similar to selecting donors with replacement.

Nonresponse Mechanism	Bias of $\hat{P}_1$.	Rel. Bias of $\hat{P}_1$.	Bias of $\hat{P}_2$.	Rel. Bias of $\hat{P}_2$.
uniform	$0.73*10^{-4}$ ($0.84*10^{-4}$)	0.13 %	$2.59*10^{-4}$ ($1.4*10^{-4}$)	0.13 %
uniform within classes	$-5.87*10^{-4}$ ($0.83*10^{-4}$)*	-1.04 %	$11.7*10^{-4}$ ($1.2*10^{-4}$)*	0.63 %
MAR (Model A1)	$10.92*10^{-4}$ ($0.75*10^{-4}$)*	1.93 %	$1.8*10^{-4}$ ($1.3*10^{-4}$)	1.0 %
MAR (Model A2)	$2.69*10^{-4}$ ($0.69*10^{-4}$)*	0.47 %	$25.3*10^{-4}$ ($1.5*10^{-4}$)*	1.39 %
MAR (Model A3)	$2.80*10^{-4}$ ($0.68*10^{-4}$)*	0.49 %	$26.2*10^{-4}$ ($1.5*10^{-4}$)*	1.41 %
MAR (Model B)	$-6.5*10^{-4}$ ($0.68*10^{-4}$)*	-1.16 %	$37.3*10^{-4}$ ($1.6*10^{-4}$)*	2.01 %
MAR (Model C)	$2.2*10^{-4}$ ($0.69*10^{-4}$)*	0.39 %	$25.3*10^{-4}$ ($1.4*10^{-4}$)*	1.35 %

Table 3.5: Simulation results for the point estimators $\hat{P}_1$. and $\hat{P}_2$. under hot deck imputation within classes selecting donors with replacement. The standard error of the bias is given in brackets. (A star (*) indicates that the bias is significantly different from zero on a 95% significance level.)

Nonresponse Mechanism	Bias of $\hat{P}_1$.	Relative Bias of $\hat{P}_1$.	Bias of $\hat{P}_2$.	Rel. Bias of $\hat{P}_2$.
uniform	$1.70*10^{-4}$ ($083*10^{-4}$)	0.31 %	$-9.8*10^{-5}$ ($1.5*10^{-4}$)	-0.05 %
uniform within classes	$-4.8*10^{-4}$ ($0.82*10^{-4}$)*	-0.85 %	$8.9*10^{-4}$ ($1.2*10^{-4}$)*	0.48 %
MAR (Model A1)	$12.12*10^{-4}$ ($0.77*10^{-4}$)*	2.15 %	$18.1*10^{-4}$ ($1.2*10^{-4}$)*	0.98 %
MAR (Model A2)	$2.04*10^{-4}$ ($0.70*10^{-4}$)*	0.36 %	$25.3*10^{-4}$ ($1.2*10^{-4}$)*	1.34 %
MAR (Model A3)	$2.54*10^{-4}$ ($0.69*10^{-4}$)*	0.45 %	$28.1*10^{-4}$ ($1.2*10^{-4}$)*	1.55 %
MAR (Model B)	$-7.7*10^{-4}$ ($0.71*10^{-4}$)*	-1.36 %	$37.2*10^{-4}$ ($1.2*10^{-4}$)*	1.97 %
MAR (Model C)	$2.4*10^{-4}$ ($0.68*10^{-4}$)*	0.44 %	$25.1*10^{-4}$ ($1.3*10^{-4}$)*	1.33 %

Table 3.6: Simulation results for point estimators $\hat{P}_1$. and $\hat{P}_2$. under hot deck imputation within classes selecting donors without replacement. (A star (*) indicates that the bias is significantly different from zero on a 95% significance level.)

3.4 Evaluation under Alternative Assumptions

When evaluating imputation methods it is of interest to investigate the robustness against model misspecification. In the following the use of different imputation models and the impact of different choices of imputation classes are investigated. In addition, hot deck imputation within classes is applied under different specifications of the imputation method. More on the analysis of underlying assumptions and alternative specifications of predictive mean matching imputation,

such as the number of repeated imputations, stochastic and deterministic imputations, can be found in chapter 5. In chapter 6 the imputation method is analysed under a nonignorable nonresponse mechanism which reflects departure from the assumption of MAR.

3.4.1 Sensitivity Analysis of Misspecification of the Imputation Model

When using imputation methods based on an imputation model it is of interest to analyse the impact of different specifications of the model. Parametric methods, such as random or deterministic regression imputation, can be affected by model misspecification, whereas semi- or nonparametric methods, such as random hot deck and nearest neighbour imputation, are much less prone to violations of the underlying model (Schenker and Taylor, 1996; Chen and Shao, 2000). In the following the previously described simulation study is used to assess the impact of different regression models in the imputation process. For each sample $s^{(a)}$ the variable $\ln(Y)$ is generated according to model (3.17), based on the variables specified in section 2.2.4. The imputation is carried out based on different linear regression models to investigate the robustness against departure of the model that underlies the data. Three different nonresponse mechanisms are used, uniform, uniform within classes and MAR nonresponse based on model A3. For simplicity we only analyse effects on estimator $\hat{P}_1$..

The results are presented in tables 3.7-3.9. A star (*) indicates that the bias is significantly different from zero on a 95% significance level. As expected, under uniform nonresponse the imputation method is robust against model misspecification and apart from one case no significant bias is found. For uniform within classes and MAR nonresponse, imputation models that only include $\ln(X)$, $\ln(X)+\ln(X)^2$ or all main effects but exclude $\ln(X)$ show a non-negligible effect on the bias of the estimator with a significant relative bias between 7% and 15%. Overall the effect in the case of MAR nonresponse seems stronger with larger biases under misspecification. A high relative bias of 15% was found for the model including significant variables apart from $\ln(X)$ under uniform within classes nonresponse. All models including $\ln(X)$ and at least some of the other significant variables, with or without non-linear terms, perform well with nonsignificant or negligible bias under all response mechanisms. This result might be expected since the derived variable is highly significant in predicting hourly pay. However, this variable on its own is not

necessarily sufficient and additional variables are needed to improve the predictive power of the imputation model.

We can conclude that the variable $\ln(X)$ seems to be important for the performance of the imputation method under the conditions considered here. Excluding this variable has a non-negligible impact on the bias of the point estimator. The imputation method is sensitive to violent model misspecification, which have a negative impact on the performance of the point estimator but seems robust against minor model misspecification. Very good results are obtained for all models that include the variable $\ln(X)$ and at least some of the other significant explanatory variables, with or without nonlinear terms and interactions. Careful modelling of the imputation model seems therefore important for the performance of the imputation method.

Under uniform nonresponse:

Explanatory Variables used in Imputation Model	Bias of $\hat{P}_1$.	Rel. Bias of $\hat{P}_1$.
$\ln(X)$	$1.6*10^{-4}$ $(2.8*10^{-5})^*$	0.29 %
$\ln(X)+\ln(X)^2$	$1.1*10^{-4}$ $(8.9*10^{-5})$	0.19 %
all main effects apart from $\ln(X)$	$1.3*10^{-4}$ $(8.6*10^{-5})$	0.23 %
simpler model including $\ln(X)$ [1]	$1.2*10^{-4}$ $(8.1*10^{-5})$	0.22 %
all main effects including $\ln(X)$ but no nonlinear terms and no interactions	$1.1*10^{-4}$ $(8.4*10^{-5})$	0.19 %
all main effects, nonlinear terms and interactions	$0.7*10^{-4}$ $(8.4*10^{-5})$	0.13 %

Table 3.7: Bias and relative bias of the point estimator $\hat{P}_1$. for different imputation models under hot deck imputation within classes, selecting donors with replacement, under uniform nonresponse.

Under uniform within classes nonresponse:

Explanatory Variables used in Imputation Model	Bias of $\hat{P}_1$.	Rel. Bias of $\hat{P}_1$.
$\ln(X)$	$4.1*10^{-3}$ $(1.4*10^{-4})^*$	7.33 %
$\ln(X)+\ln(X)^2$	$4.3*10^{-3}$ $(1.5*10^{-4})^*$	7.75 %
all main effects apart from $\ln(X)$	$8.6*10^{-3}$ $(2.8*10^{-4})^*$	15.21 %
simpler model including $\ln(X)$	$1.1*10^{-3}$ $(0.7*10^{-4})^*$	2.05 %
all main effects including $\ln(X)$ but no nonlinear terms and no interactions	$0.3*10^{-3}$ $(0.7*10^{-4})^*$	0.59 %
all main effects, nonlinear terms and interactions	$0.3*10^{-3}$ $(0.7*10^{-4})^*$	0.49 %

Table 3.8: Bias and relative bias of the point estimator $\hat{P}_1$. for different imputation models under hot deck imputation within classes, selecting donors with replacement, under uniform within classes nonresponse.

Under MAR nonresponse:

Explanatory Variables used in Imputation Model	Bias of $\hat{P}_1$.	Rel. Bias of $\hat{P}_1$.
$\ln(X)$	$0.6*10^{-3}$ $(2.1*10^{-4})^*$	10.68 %
$\ln(X)+\ln(X)^2$	$0.6*10^{-3}$ $(2.1*10^{-4})^*$	10.68 %
all main effects apart from $\ln(X)$	$5.1*10^{-3}$ $(1.7*10^{-4})^*$	9.10 %
simpler model including $\ln(X)$	$1.7*10^{-3}$ $(0.9*10^{-4})^*$	3.03 %
all main effects including $\ln(X)$ but no nonlinear terms and no interactions	$0.1*10^{-3}$ $(0.7*10^{-4})$	0.21 %
all main effects, nonlinear terms and interactions	$0.3*10^{-3}$ $(0.7*10^{-4})^*$	0.49 %

Table 3.9: Bias and relative bias of the point estimator $\hat{P}_1$. for different imputation models under hot deck imputation within classes, selecting donors with replacement, under MAR nonresponse.

[1] The simpler model includes: $\ln(X)$, $\ln(X)^2$, major occupation group (SOCcat), qualification (Qcat), AGE, AGE^2, industry section (INCcat), regions (REGcat) and gender)

3.4.2 Robustness Against the Choice of Imputation Classes

A potential drawback of the hot deck imputation within classes is the choice of imputation classes. The classes are defined by the range of the predicted values of the linear regression model. However, the borders are effectively chosen arbitrarily. There is some grounds to believe that the estimator $\hat{P}_{\cdot}$ may depend on the choice of imputation classes. In particular, this is of interest, since the level of the NMW increases over time and this level might move towards a boundary of a class. Since the imputation is carried out within classes this could affect the estimation of the parameter of interest. Changing the boundaries of the imputation classes with each increase of the NMW, such that the level of the NMW is well covered by one of the classes, may affect comparability of the estimates for different quarters of the LFS. In the following the effect of imputation classes is analysed using a simulation study, as described in section 3.3. For simplicity only 22+ year olds are taken into account. The boundaries of the response and imputation classes are held fixed. The effect of different levels of the NMW, denoted t, are investigated, where

a.) $t = \ln(£4.00) = 1.38$, such that t is close to the upper boundary of the first class,

b.) $t = \ln(£4.10) = 1.41$, such that t is close to the lower boundary of the second class and

c.) $t = \ln(£4.95) = 1.59$, such that t is close to the upper boundary of the second class.

Level of NMW	Nonresponse Mechanism	Bias of $\hat{P}_{1.}$	Rel. Bias of $\hat{P}_{1.}$	Bias of $\hat{P}_{2.}$	Rel. Bias of $\hat{P}_{2.}$
t= ln(£4.00)	unif. in classes	$9.2*10^{-4}$ $(0.93*10^{-4})^*$	0.93 %	$4.9*10^{-4}$ $(1.1*10^{-4})^*$	0.34 %
	MAR (Model A3)	$12.1*10^{-4}$ $(0.97*10^{-4})^*$	1.30 %	$15.2*10^{-4}$ $(1.2*10^{-4})^*$	1.04 %
t= ln(£4.10)	unif. in classes	$4.5*10^{-4}$ $(0.95*10^{-4})^*$	0.40 %	$9.0*10^{-4}$ $(1.1*10^{-4})^*$	0.69 %
	MAR (Model A3)	$9.7*10^{-4}$ $(0.94*10^{-4})^*$	0.86 %	$18.2*10^{-4}$ $(1.2*10^{-4})^*$	1.42 %
t= ln(£4.95)	unif. in classes	$17.2*10^{-4}$ $(1.3*10^{-4})^*$	0.73 %	NA	NA
	MAR (Model A3)	$31.0*10^{-4}$ $(1.6*10^{-4})^*$	1.35 %	NA	NA

Table 3.10: Bias and relative bias of $\hat{P}_{1.}$ and $\hat{P}_{2.}$ under uniform within classes and MAR nonresponse using different levels of the NMW, selecting donors with replacement.

Table 3.10 presents bias and relative bias of $\hat{P}_{1.}$ and $\hat{P}_{2.}$ under uniform within classes and MAR nonresponse using different levels of the NMW. We can see that all results are significantly biased. However, the (estimated) bias is small and the relative bias for both point estimators are well below 2%. In the case of MAR nonresponse the relative bias is with 1.30%, 0.86% and 1.35% higher for all three levels of t than for the original level of the NMW, where t=ln(£3.60), with a relative bias of 0.49%. Also for uniform within classes nonresponse the relative bias for t=ln(£4.00) is with 0.93% higher than for the original level of t. Calculating the 95% confidence intervals for the bias of $\hat{P}_{1.}$ it was found that the intervals for all three levels of t were further away from zero and non-overlapping with the interval for t=ln(£3.60) in the case of MAR nonresponse. This result gives an indication that certain specifications of the imputation classes might lead to a (small) increase in the bias of the point estimators.

Another possibility to assess the impact of the choice of imputation classes on the estimator is to define the boundaries of the imputation classes differently, resulting in an increase or a decrease of the total number of imputation classes. We analysed the effects on the estimators by narrowing the width of the imputation classes. Instead of 7 we defined 28 imputation classes. The results for the bias and relative bias under 28 classes are presented in table 3.11. All biases are significant. The results are found to be very similar to the results when using 7 classes, as given in table 3.5. An analysis where the boundaries of the response and imputation classes under uniform within classes nonresponse are defined differently was not carried out.

Nonresponse Mechanism	Bias of $\hat{P}_1$.	Rel. Bias of $\hat{P}_1$.	Bias of $\hat{P}_2$.	Rel. Bias of $\hat{P}_2$.
unif. in classes	$-6.01*10^{-4}$ $(0.81*10^{-4})^*$	-1.22 %	$10.1*10^{-4}$ $(1.3*10^{-4})^*$	0.52 %
MAR	$2.41*10^{-4}$ $(0.68*10^{-4})^*$	0.41 %	$28.1*10^{-4}$ $(1.2*10^{-4})^*$	1.65 %

Table 3.11: Bias and relative bias of $\hat{P}_1$. and $\hat{P}_2$. under uniform within classes and MAR nonresponse using 28 imputation classes.

In summary, the results suggest that the hot deck imputation method within classes seems to be reasonably robust against different specifications of the imputation classes. An indication is found that the method might be sensitive to an increase in the level of the NMW. An increase of the level of the NMW and therefore a shift of this threshold towards a boundary of an imputation class may cause an increase in the bias of the point estimator. However, the effect is expected to be small.

3.4.3 Analysis of the Effects of Different Specifications of the Imputation Method Applied to LFS Data

The described hot deck imputation method has been implemented by ONS. In this section the method as used by ONS in practice, in addition to using various alternative specifications, are

applied to LFS data. The aim is to assess the robustness of the proposed method under different specifications, such as the use of different imputation models.

The covariates in the imputation model used by ONS include the derived variable, occupation, qualification, industry, region and other variables.[2] The model initially did not include squared terms or interactions. In total 16 imputation classes were used, each 50 pence wide, with the top class containing employees earning £15 per hour or more. Employees in the group of professionals and associate professionals were treated separately in the imputation process, since they showed a different behaviour in the residuals obtained from the imputation model. Therefore donor values for a nonrespondent belonging to this group also had to be chosen from this group. Donor values were selected without replacement to avoid using donors disproportionately often and to minimise the variance inflation effect in the presence of imputation. In addition, it was decided to exclude outliers from the imputation process such that outlying cases would not have an influencing effect on the imputed estimates. Outliers were defined as those cases where the residual $\varepsilon_i = \ln(y_i) - \ln(\hat{y}_{regi})$ fell outside the 1st or 99th percentile of the distribution of the residuals. These cases were excluded from the set of potential donors.

Table 3.12 presents the estimates for $\hat{P}_1.$ and $\hat{P}_2.$ under the hot deck imputation method as carried out by ONS under the above specifications. In the following different specifications of the method and possible effects on the resulting estimates are discussed keeping all other specifications fixed. As can be seen from table 3.12 there is an indication that different imputation models can have an effect on the estimates, using for comparison a more detailed imputation model including squared terms, interactions and a greater number of variables, such as variables related to work pattern and payment methods, and a simpler model based on a much smaller number of variables and excluding nonlinear terms.[3] With increasing complexity of the imputation model a reduction in the estimates for both point estimators is observed. This might reflect a possible departure from the MAR assumption for simpler imputation models. Treating the group

[2] The variables included in the ONS imputation model are:
ln(derived), whether or not part time, occupation, head of household, married, qualifications, pay period less than weekly, months continuously employed, size of workplace, industry section and region.

[3] The variables included in the more detailed model are: all variables as before plus gender (female), age, age squared, two youth indicator variables, temporary work, ever worked overtime, pay period less than monthly, whether additions to basic pay, whether last pay same as usual. The variable head of household was dropped.
The variables included in the simpler model are: ln(derived), age, gender, industry sector, paid less than weekly, part-time, qualifications and major occupation group.

of professional and associate professionals differently in the imputation process had little impact. A different specification of the width of the imputation classes, using classes that are only 25 pence wide, which increased the total number of classes, did not seem to have a great impact on the overall estimates. This coincides with the findings in section 3.4.2. Sampling donors with replacement instead of without replacement did not affect the estimates. The exclusion of outliers, however, showed an effect on the estimates leading to a reduction in the proportion of low paid employees, since cases were excluded with large negative residuals such that the observed value for Y was small. We conclude that the use of different imputation models as well as the treatment of outliers can have an effect on the overall estimates, whereas the hot deck imputation method seems to be robust against all other specifications discussed here. Note that all estimates reported here are based on a stochastic imputation method and that a repetition of the method under the same specifications can lead to slight variations in the results.

Imputation Method	$\hat{P}_1$. **(weighted)**	$\hat{P}_2$. **(weighted)**
Hot deck imputation within classes under specifications as carried out by ONS	1.53	27.40
Modifications to Method		
More detailed imputation model	1.31	26.37
Simpler imputation model	1.60	27.41
Professionals not treated separately	1.51	27.09
25p bands	1.51	27.31
With replacement	1.52	27.44
Outliers not excluded	2.00	28.79

Table 3.12: Estimates for both point estimators $\hat{P}_1$. and $\hat{P}_2$. (weighted) under hot deck imputation method within classes as carried out by ONS and under modifications of the method for age group 18+, June-August 1999.

3.5 Conclusion

This chapter evaluated hot deck imputation within classes with respect to the point estimator of interest. The properties of the imputation method were analysed both theoretically and empirically. It was shown that under certain conditions the point estimator is approximately unbiased, both unweighted and weighted, assuming fixed survey weights. A simulation study based on LFS data was carried out to evaluate the imputation method empirically. Under the conditions of the simulation study both point estimators investigated were found to be approximately unbiased. The imputation method was also analysed under misspecification of the imputation model. It was found that the method is robust to minor model misspecification, however, sensitive to violent misspecification. There are some grounds to believe that the performance of the hot deck method may depend on the choice of classes. In particular, if the imputation classes do not correspond to the response classes in the population bias may be introduced. An increase in the bias was observed for some misspecifications, however, in the investigations carried out here the bias was found to be small. In chapters 5 and 6 further investigation of the impact of different choices of imputation classes is carried out.

Chapter 4

Variance and Variance Estimation for Hot Deck Imputation Within Classes

Valid variance estimation procedures for point estimators that are based on imputed values are needed since standard variance formulae are invalid when imputation has been used. Applying a naive variance formula could lead to considerable underestimation of the true variance. One approach of solving this problem has been proposed by Rubin (1987) using proper multiple imputation (MI) in a Bayesian framework (see chapter 1). However, particularly in governmental and statistical agencies, imputation methods other than proper multiple imputation are commonly used, such as hot deck or regression imputation. Therefore, there is a need to develop valid variance estimation in the presence of such imputation methods.

The aim is to derive a valid variance estimator for the hot deck procedure described in chapters 2 and 3. Section 4.1 gives an overview over possible approaches, including a two-phase approach, a model-assisted approach and replication methods. In section 4.2 variance estimation for the hot deck imputation method not taking into account LFS survey weights is developed. In section 4.3 weighting is taken into account. To evaluate the performances of several variance estimators a simulation study is carried out, described in section 4.4. In section 4.5 a proper MI method is implemented based on the approximate Bayesian Bootstrap. The performance of Rubin's MI formula is investigated under this approach. For simplicity stratification is not taken into account.

Stratification has been addressed for example in Chen and Shao (2001) and Rao and Shao (1992). Also clustering of employees into households is not taken into account. If clustering were considered the imputation method may need to be carried out within clusters. Some concluding remarks are given in section 4.6.

4.1 Variance and Variance Estimation in the Presence of Imputation

This section briefly reviews different approaches to variance estimation of a point estimator when single value imputation, in particular hot deck and regression imputation, has been used. The problem of variance estimation depends on the response mechanism, the imputation method used, the sampling design and the type of point estimator. Different assumptions about the response mechanism require different approaches, for example a design based approach without stating a model or a model-assisted approach. In the following we denote the expectation and the variance according to the design as E_D and var_D and according to the response mechanism as E_R and var_R, using the notation as introduced in chapter 3. Instead of writing $E_D E_R(.|s)$, we write $E_{DR}(.)$.

4.1.1 Mean Square Error of an Imputed Estimator

In the following we refer to the parameter of interest θ, which is estimated by $\hat{\theta}$ based on sample data in the case of full response and by $\hat{\theta}.$ in the presence of imputed data using notation as in chapter 1. The total error of the imputed estimator $\hat{\theta}.$ can be divided into *sampling error*, $\hat{\theta}-\theta$, and *imputation error*, $\hat{\theta}.-\hat{\theta}$, such that

$$\hat{\theta}.-\theta=(\hat{\theta}-\theta)+(\hat{\theta}.-\hat{\theta}). \tag{4.1}$$

We assume that the chosen point estimator $\hat{\theta}$ is approximately design-unbiased for the parameter θ. In general we may use the mean square error (MSE) as a measure of the quality of the imputed estimator. The MSE of $\hat{\theta}.$ with respect to D and R is given by (see also Lee, Rancourt and Särndal, 2000; Deville and Särndal, 1994)

$$MSE_{DR}(\hat{\theta}.) = E_{DR}((\hat{\theta}.-\theta)^2) = E_{DR}(((\hat{\theta}.-\hat{\theta})+(\hat{\theta}-\theta))^2)$$

$$= E_{DR}((\hat{\theta}-\theta)^2) + E_{DR}((\hat{\theta}.-\hat{\theta})^2 + 2(\hat{\theta}.-\hat{\theta})(\hat{\theta}-\theta))$$

$$= \mathrm{var}_D(\hat{\theta}) + E_D \mathrm{var}_R(\hat{\theta}.|s) + E_D((E_R(\hat{\theta}.|s)-\hat{\theta})^2) + 2\mathrm{cov}_D(\hat{\theta}, E_R(\hat{\theta}.|s)-\hat{\theta})$$

$$= V_{sam} + V_{res} + E_D(B_c^2) + 2\mathrm{cov}_D(\hat{\theta}, B_c), \tag{4.2}$$

where $B_c = E_R(\hat{\theta}.|s) - \hat{\theta}$ is the conditional bias, V_{sam} is the standard variance according to the sampling design and $V_{res} = E_D \mathrm{var}_R(\hat{\theta}.|s)$ is the variance caused by nonresponse followed by imputation, which also takes into account variation of the number of respondents. Lee, Rancourt and Särndal (2000) denote

$$V_{cres} = \mathrm{var}_R(\hat{\theta}.|s) \tag{4.3}$$

as the conditional variance. The term $V_{res} + E_D(B_c^2) + 2\mathrm{cov}_D(\hat{\theta}, B_c)$ measures the increase in the MSE caused by imputation due to nonresponse in the sample. The term $E_D(B_c^2) + 2\mathrm{cov}_D(\hat{\theta}, B_c)$ represents the bias of $\hat{\theta}.$, where in particular $E_D(B_c^2)$ can be large. If we assume that the imputed estimator is unbiased, i.e. $B_c = E_R(\hat{\theta}.|s) - \hat{\theta} = 0$, the MSE is equal to the total variance of $\hat{\theta}.$:

$$MSE_{DR}(\hat{\theta}.) = V_{tot} = V_{sam} + V_{res}. \tag{4.4}$$

We see that the total variance consists of two components, the variance according to sampling and the variance caused by nonresponse followed by imputation. When estimating the variance in the presence of imputation it can be useful to estimate these two components separately (Deville and Särndal, 1994; Lee, Rancourt and Särndal, 2000). Note that it is possible to incorporate explicitly a term for the variability due to random imputation. This will be done in sections 4.2 and 4.3.

In the following section we consider different approaches to variance estimation, in particular differences between design-based (section 4.1.2) and model-assisted approaches (section 4.1.3). Each method is based on certain assumptions about the nonresponse mechanism. Alternative methods are jackknife and bootstrap estimation, which are described in section 4.1.4.

4.1.2 Variance Estimation Using a Two Phase Method

The joint distribution *DR* may be viewed as two-phase sampling, where in the first phase the sample is selected according to *D* and in the second phase the respondent set is selected from the first phase sample according to *R*. In order to proceed we need to make assumptions about the response mechanism *R*, which is usually unknown. A common approach is to assume either uniform response or uniform response within classes. This method has a natural design-based interpretation and has been shown to be robust to the misspecification of the imputation model (Lee, Rancourt and Särndal, 1994; Rao and Sitter, 1995; Shao and Steel, 1999). We again assume that $\hat{\theta}.$ is approximately design unbiased for θ. Using (4.4) we aim to find (approximately) unbiased estimators of V_{sam} and V_{res}, such that

$$E_{DR}(\hat{V}_{sam}) \doteq V_{sam} \text{ and } E_{DR}(\hat{V}_{res}) \doteq V_{res}. \tag{4.5}$$

The latter condition holds if we find an (approximately) unbiased estimator of $V_{c\,res}$, the conditional variance defined in (4.3), such that $E_R(\hat{V}_{c\,res}) \doteq V_{c\,res} = \mathrm{var}_R(\hat{\theta}.|\,s)$ for all samples *s*. It follows

$$\hat{V}_{tot} = \hat{V}_{sam} + \hat{V}_{c\,res} \tag{4.6}$$

such that $E_{DR}(\hat{V}_{tot}) \doteq V_{tot}$. The advantage of this approach is that it only requires two distributions, the sampling design and the response mechanism. However, this method relies on the assumptions of the response mechanism. If the assumptions made are not true, the variance estimator is biased. It should be mentioned that under a stochastic imputation method and assuming a uniform response mechanism within imputation classes, an approximately unbiased estimator of V_{sam} can usually be found by applying the naive variance estimator on the imputed data. For the imputation method discussed in chapter 2 and 3 this will be shown in section 4.2.2. Under deterministic imputation, however, the naive variance estimator often leads to underestimation of the sampling variance (Kovar and Chen, 1994). Several examples of how to obtain the estimators $\hat{V}_{sam}$ and $\hat{V}_{c\,res}$ are discussed in Lee, Rancourt and Särndal (2000).

Rao and Shao (1992; see also Oh and Scheuren, 1983) derive an estimator of the total variance for hot deck imputation assuming simple random sampling and a uniform response mechanism. The

parameter of interest is the population mean of the variable Y, estimated by $\hat{\theta}_{\cdot} = \bar{y}_{\cdot}$. Denote $s_{\cdot}^2$ the imputed sample variance, i.e. the naive variance formula treating imputed values as if they were real values. Since $s_{\cdot}^2/n$ is a biased variance estimator for $\bar{y}_{\cdot}$. Rao and Shao (1992) propose a correct estimator of the variance. The approximately unbiased variance estimator ignoring the finite population correction is given by

$$\hat{V}_{tot}(\bar{y}_{\cdot}) = \left(1+(\frac{n_r}{n})(\frac{n_{\bar{r}}}{n})\right)\frac{s_{\cdot}^2}{n_r} = (\frac{n}{n_r}+\frac{n_{\bar{r}}}{n})\frac{s_{\cdot}^2}{n}, \tag{4.7}$$

since the sample of respondents is a simple random sample of size n_r from the population of size N. Approximate estimators of the two components V_{sam} and V_{res} as in (4.5) are

$$\hat{V}_{sam} = \frac{s_{\cdot}^2}{n} \text{ and } \hat{V}_{res} = \frac{s_{\cdot}^2}{n}\left(\frac{n_{\bar{r}}}{n_r}+\frac{n_{\bar{r}}}{n}\right), \tag{4.8}$$

which together give $\hat{V}_{tot}(\bar{y}_{\cdot})$ in (4.7). Note that in this particular example $s_{\cdot}^2/n$ is an adequate estimator for V_{sam}. Rao and Shao (1992), however, do not extend the formula to uniform within classes nonresponse.

4.1.3 Variance Estimation Using a Model-Assisted Approach

When applying imputation an imputation model is used either implicitly or explicitly. We denote the imputation model as ξ, which is in a general context given by

$$y_i = \mu_i\beta + \varepsilon_i, \tag{4.9}$$

where μ_i is a row-vector of functions of the covariates W and β is a vector of regression coefficients, $E_\xi(\varepsilon_i)=0$, $E_\xi(\varepsilon_i^2)=c_i\sigma^2$, where c_i are suitably defined constants and $E_\xi(\varepsilon_i\varepsilon_k)=0$ if $i \neq k$ and E_ξ denotes the expectation with respect to the model. Many imputation methods involve the estimation of the parameters in the model from the respondents in the sample. Most imputation methods can be expressed as a form of regression imputation (Kalton and Kasprzyk, 1986), such that

$$\hat{y}_j = \mu_j \hat{\beta}_r + \hat{\varepsilon}_j \quad \forall j \in \bar{r}, \tag{4.10}$$

where $\hat{y}_j$ is the imputed value for Y and $\hat{\beta}_r$ is a vector of estimated coefficients based on respondent data (Särndal, Swensson and Wretman, 1992; Deville and Särndal, 1994; Lee, Rancourt and Särndal, 2000; Rao, 2001). Under a random regression imputation method the residuals $\hat{\varepsilon}_j$ are selected at random, e.g. following a normal distribution. For deterministic imputation one sets $\hat{\varepsilon}_j$ =0 for all $j \in \bar{r}$. Hot deck imputation can be expressed in this form by including dummy auxiliary variables to represent the classes (Lee, Rancourt and Särndal, 2000).

In this section the model-assisted approach, as defined in Deville and Särndal (1994), is described. According to Deville and Särndal this approach requires that the imputation model given in (4.9) is stated explicitly and the derivation of a variance formula is based on the model assumptions. Using this technique, the distribution with respect to the imputation model ξ in addition to the distribution with respect to the sampling design and the nonresponse are taken into account. This method allows to make a weaker assumption about the nonresponse mechanism, namely MAR, than before in the two-phase sampling approach. The drawback of this method is that it is sensitive to the imputation model assumptions. This approach will not be used in the further analysis, however, it is described because of its relevance.

In the following MAR nonresponse is assumed. It is also assumed that the estimator $\hat{\theta}.$ is an approximately ξDR-unbiased estimator of θ. This assumption will generally follow if the imputation model is correct. The expectation and variance referring to the imputation model ξ are denoted E_ξ and var_ξ respectively. Let us consider the ξDR-variance

$$\mathrm{var}_{\xi DR}(\hat{\theta}.) = E_\xi E_D E_R((\hat{\theta}. - \theta)^2) = E_\xi \mathrm{var}_{DR}(\hat{\theta}.). \tag{4.11}$$

Note that

$$E_\xi \mathrm{var}_{DR}(\hat{\theta}.) = E_\xi V_{tot} = E_\xi V_{sam} + E_\xi V_{res}. \tag{4.12}$$

It can be seen that similar to the two phase approach the two components of the total variance can be estimated separately. The estimators $\hat{V}_{sam}$ and $\hat{V}_{res}$ must fulfil the following conditions

$$E_\xi(E_D E_R(\hat{V}_{sam}) - V_{sam}) \doteq 0 \text{ and } E_\xi(E_D E_R(\hat{V}_{res}) - V_{res}) \doteq 0, \tag{4.13}$$

such that $E_{\xi}(E_D E_R(\hat{V}_{tot}) - V_{tot}) \doteq 0$. The latter condition in (4.13) reduces to $E_{\xi}(E_R(\hat{V}_{c\,res}) - V_{c\,res}) \doteq 0$, where $V_{c\,res}$ is defined in (4.3). Note that conditions (4.13) are the model equivalent conditions to (4.5). It follows as before

$$\hat{V}_{tot} = \hat{V}_{sam} + \hat{V}_{c\,res}. \tag{4.14}$$

We therefore seek estimators of the two components, i.e. we seek $\hat{V}_{sam}$ and $\hat{V}_{cres}$. Examples of how to find such approximately unbiased estimators of the two components can be found in Deville and Särndal (1994) and in Lee, Rancourt and Särndal, (1995 and 2000). Note that the computed variance component estimates obtained under the model assisted approach are usually different from the two phase approach (Deville and Särndal, 1994).

The separate estimation of the variance components is often useful from the survey design point of view. Both approaches, the two-phase approach and the model-assisted approach, allow this, whereas methods such as jackknife and bootstrap do not. When choosing an imputation method the objective should be to minimise the total variance. If the imputation variance makes up a considerable part of the total variance then the focus should be on reducing this impact by modifying the imputation method. If a stochastic single value imputation is used multiple times (improper multiple imputation) with the aim of reducing variability due to imputation then the single value variance estimator with modification can be applied (Lee, Rancourt and Särndal, 2000; Rao, 2001).

4.1.4 Variance Estimation Using Resampling Methods

An alternative way of estimating the variance of an estimator in the presence of imputation is to use resampling methods such as the bootstrap and the jackknife (Little and Rubin, 2002). These two approaches, designed to obtain variance estimates without having to derive a closed expression (Wolter, 1985), will be briefly discussed in the following. Let s be a sample of size n, $s = \{i : i, ..., n\}$, of independent observations including missing data. The Bootstrap method requires the generation of Bootstrap samples $s^{(b)}$, $b = 1, ..., B$ each of size n by resampling units by simple random sampling with replacement from the original unimputed sample s. The imputation method is carried out in each bootstrap sample $s^{(b)}$. Let $\hat{\theta}^{(b)}_{.}$ be the imputed estimator in

bootstrap sample $s^{(b)}$. Then the bootstrap estimate of θ is the average of the imputed bootstrap estimates

$$\hat{\theta}_{.} = \hat{\theta}_{.boot} = \frac{1}{B}\sum_{b=1}^{B}\hat{\theta}_{.}^{(b)} . \tag{4.15}$$

A consistent estimate of the variance of $\hat{\theta}_{.}$ or $\hat{\theta}_{.boot}$ is

$$\hat{V}_{boot} = \frac{1}{B-1}\sum_{b=1}^{B}(\hat{\theta}_{.}^{(b)} - \hat{\theta}_{.boot})^2 . \tag{4.16}$$

It is important that the imputation method needs to be applied to the bootstrap samples in total B times and not to the original sample s (Shao and Sitter, 1996). Shao and Sitter (1996) also show the consistency of the bootstrap variance estimator for commonly used imputation methods and different types of statistics under complex sample designs. A disadvantage of the method is that it is computationally intensive. Alternatively, instead of applying the imputation to each replicate, it is possible to use an adjustment approach to the standard bootstrap adjusting imputed values. However, the adjustment approach is not suitable for quantile estimation (Lee, Rancourt and Särndal, 2000). The approach presented here requires large samples. With smaller samples the imputation method may need to be modified for the bootstrap samples. (Shao and Sitter, 1996; Lee, Rancourt and Särndal, 2000; Little and Rubin, 2002).

The jackknife is similar to the bootstrap but requires the successive dropping of units from the original sample (Little and Rubin, 2002). Let s be as above. In the classical case with no complex design the jackknife sample is denoted $s^{(\setminus i)}$ of size $n-1$, obtained from sample s by dropping the ith unit. The imputation procedure is carried out within each jackknife sample. Let $\hat{\theta}_{.}$ be a consistent estimate of θ obtained by imputing the missing values in the original sample s and $\hat{\theta}_{.}^{(\setminus i)}$ be the imputed estimator computed based on the jackknife sample. Then the pseudovalue

$$\tilde{\theta}_{._j} = n\hat{\theta}_{.} - (n-1)\hat{\theta}_{.}^{(\setminus j)} \tag{4.17}$$

can be computed and the jackknife point estimator is given by

$$\hat{\theta}_{.jack} = \frac{1}{n}\sum_{j=1}^{n}\tilde{\theta}_{.j} = \hat{\theta}_{.} + (n-1)(\hat{\theta}_{.} - \bar{\theta}_{.}) \tag{4.18}$$

where $\bar{\theta}_{.} = \frac{1}{n}\sum_{i=1}^{n}\hat{\theta}_{.}^{(\setminus i)}$. The jackknife estimate of the variance of $\hat{\theta}_{.}$ or $\hat{\theta}_{.jack}$ is

$$\hat{V}_{jack} = \frac{1}{n(n-1)}\sum_{i=1}^{n}(\tilde{\theta}_{.i} - \hat{\theta}_{.jack})^2 = \frac{n-1}{n}\sum_{i=1}^{n}(\hat{\theta}_{.}^{(\setminus i)} - \bar{\theta}_{.})^2, \tag{4.19}$$

which is a consistent estimate of the variance of the imputed point estimator $\hat{\theta}_{.}$ (Little and Rubin, 2002, p. 83). However, this depends on the type of imputation. The approach is valid for deterministic imputation, however for random imputation the method leads to overestimation of the variance. Rao and Shao (1992) report that in the case of hot deck imputation the jackknife method using re-imputation leads to overestimation of the true variance if n is large and show this result theoretically (see also Lee, Rancourt and Särndal, 2000). They propose a technique that adjusts the imputed values to correct the jackknife variance estimator in the case of imputed data. The imputation is carried out in the original sample s. The basic principle of the adjustment in each jackknife sample is that whenever a responding unit is deleted each of the imputed values in the jackknife sample is adjusted. The imputed values are unchanged if a nonresponding unit is deleted. The adjustments depend on the expectation with respect to the imputation procedure applied in the replicate sample (Lee, Rancourt and Särndal, 2000). Therefore under deterministic imputation the adjustment method by Rao and Shao (1992) is equivalent to the method discussed in Little and Rubin (2002). The adjusted jackknife method is design consistent as well as model unbiased. If n is large, blocks of several units can be deleted instead of single units to reduce computing time.

The jackknife can also be used in the case of a complex sampling design, for example if the sample s is a stratified cluster sample or a stratified multistage sample with imputed data. Some adjustments to the imputed values are required in the case of such a complex sampling design, which are discussed in Rao and Shao (1992) and Kovar and Chen (1994). Rao and Shao (1992) and Kovar and Chen (1994) discuss adjustments necessary for the case of hot deck imputation within classes under the assumption of uniform nonresponse within classes. Rao (1996) provides a

comprehensive account of the jackknife approach with adjustment. Rao and Sitter (1995) develop a linearized jackknife variance formula. Skinner and Rao (1993) present jackknife variance estimation for multivariate statistics under hot deck imputation. It should be noted that the jackknife cannot be applied to the case where the imputed estimator $\hat{\theta}.$ is nonsmooth, for example in the case of estimating quantiles and proportions (Shao and Wu, 1989; Shao and Sitter, 1996) and is therefore not appropriate for the application considered here.

Both the bootstrap method and the jackknife method are computer intensive, particularly if properties of the resulting estimators are analysed in a simulation study. Another disadvantage of the two methods is that no variance estimator in closed form is available. The variance formulae (4.16) and (4.19) do not allow for a separate breakdown into several variance components such as V_{sam} and V_{res}. For these reasons, jackknife and bootstrap methods are not used to obtain an estimate of the variance of the point estimator $\hat{P}..$

4.2 Variance and Variance Estimation for Hot Deck Imputation Within Classes not Taking into Account LFS Weights

In this section we derive a formula for the variance of the estimator of interest $\hat{P}.$, defined in section 2.2.5, as well as an approximately unbiased estimator of this formula taking into account imputation (I), response (R) and sampling variability (D) using a design-based approach (or three-phase approach) similarly to section 4.1.2. Initially, the complex weighting scheme of the LFS is not taken into account. The response mechanism is assumed to be uniform within classes and we consider selecting donors with replacement. In addition, we investigate variance estimation using Rubin's multiple imputation formula. This formula is designed for proper multiple imputation, however, and we find that this approach underestimates the variance for the improper imputation procedure. We therefore correct for this bias and develop an approximately unbiased modification of Rubin's estimator. The required framework and necessary definitions for the following results are given in section 3.1. Similarly to before, the notation var_{DRI} refers to the variance that takes into account imputation, response and sampling variability. Note that if there is

no nonresponse, all formulae derived here reduce to the standard variance formulae valid for full response.

4.2.1 Variance of the Estimator $\hat{P}.$

In this section we give a formula for the variance of $\hat{P}.$ under the assumption of iid observations. We show that the total variance can be divided into three components according to the variance due to imputation, response and sampling, such that

$$V_{tot} = V_{sam} + V_{imp} + V_{res}, \tag{4.20}$$

where the components are defined as

$$V_{tot} = \mathrm{var}_{DRI}(\hat{P}.),\ V_{sam} = \mathrm{var}_D E_{RI}(\hat{P}.),\ V_{imp} = E_{DR}\,\mathrm{var}_I(\hat{P}.) \text{ and } V_{res} = E_D\,\mathrm{var}_R E_I(\hat{P}.). \tag{4.21}$$

The finite population correction $(N-n)/N$ is ignored assuming N the size of the population to be infinite. Also the difference between β and $\hat{\beta}$, which determine the imputation classes, is ignored for large samples. In the following we will use Taylor approximations as described in equation (3.14). In the context used here, the Taylor series approximation requires n_{rk}, the number of respondents in B_k, and n_k, the number of units in B_k, to be reasonably large for all k.

Result 4.1

Under the assumptions of (i) equal probability sampling, (ii) uniform nonresponse within classes, (iii) n_{rk}, the number of respondents in class B_k, and n_k, the number of units in class B_k, are large for all k, and (iv) the unweighted estimator $\hat{P}.$ is obtained under the imputation method described in section 2.2.3 selecting donors by simple random sampling with replacement,
the variance of $\hat{P}.$, taking into account imputation and sampling variability and the nonresponse mechanism ignoring the finite population correction, is approximately given by

$$\mathrm{var}_{(1)DRI}(\hat{P}.) \doteq \frac{P(1-P)}{n} + \frac{1}{nN}\sum_k N_k \bar{z}_{Uk}(1-\bar{z}_{Uk})\left(\frac{1}{M}(1-\pi_k)(1-\frac{N}{N_k n\pi_k}) + (\frac{1}{\pi_k}-1)(1-\frac{N}{nN_k})\right), \tag{4.22}$$

where $\bar{z}_{Uk} = N_k^{-1} \sum_{i \in B_{Uk}} z_i$ is the proportion of low paid employees in imputation class B_{Uk} in the population, N_k the number of employees in imputation class B_{Uk} in the population and π_k the probability of response in class B_{Uk}.

Proof

The proof is given in the appendix A4.1.

As shown in this proof the total variance consists of three components

$$\text{var}_{(1)DRI}(\hat{P}.) = V_{tot} = V_{sam} + V_{imp} + V_{res} \tag{4.23}$$

of the form

$$V_{sam} \doteq \frac{P(1-P)}{n} \tag{4.24}$$

$$V_{imp} \doteq \frac{1}{MnN} \sum_k N_k \bar{z}_{Uk}(1-\bar{z}_{Uk})(1-\pi_k)\left(1-\frac{N}{N_k n \pi_k}\right) \text{ and} \tag{4.25}$$

$$V_{res} \doteq \frac{1}{nN} \sum_k N_k \bar{z}_{Uk}(1-\bar{z}_{Uk})\left(\frac{1}{\pi_k}-1\right)\left(1-\frac{N}{nN_k}\right). \tag{4.26}$$

Corollary 4.2

Under the approximations that $\frac{N}{N_k n \pi_k}$ and $\frac{N}{nN_k}$ are small, the variance of $\hat{P}.$ in result 4.1 reduces to

$$\text{var}_{(2)DRI}(\hat{P}.) \doteq \frac{P(1-P)}{n} + \frac{1}{nN} \sum_k N_k \bar{z}_{Uk}(1-\bar{z}_{Uk})\left(\frac{1}{M}(1-\pi_k) + (\frac{1}{\pi_k}-1)\right)$$

$$= \frac{P(1-P)}{n} + \frac{1}{nN} \sum_k N_k \bar{z}_{Uk}(1-\bar{z}_{Uk})\,(1-\pi_k)(\frac{1}{M}+\frac{1}{\pi_k}) \quad . \tag{4.27}$$

It follows that the total variance approximately consists of three components

$$\mathrm{var}_{(2)DRI}(\hat{P}.) = V_{tot} = V_{sam} + V_{imp} + V_{res} \tag{4.28}$$

where

$$V_{sam} \doteq \frac{P(1-P)}{n} \tag{4.29}$$

$$V_{imp} \doteq \frac{1}{MnN}\sum_k N_k \bar{z}_{Uk}(1-\bar{z}_{Uk})(1-\pi_k) \text{ and} \tag{4.30}$$

$$V_{res} \doteq \frac{1}{nN}\sum_k N_k \bar{z}_{Uk}(1-\bar{z}_{Uk})\left(\frac{1}{\pi_k}-1\right). \tag{4.31}$$

We can see that V_{sam} is the standard variance formula of a proportion. Both V_{res} and V_{imp} depend on the response rates π_k, $k = 1,...,K$. Investigating the order of the three components we see that $V_{sam} = O(\frac{1}{n})$, $V_{imp} = O(\frac{1}{Mn})$ and $V_{res} = O(\frac{1}{n})$. If M is regarded as fixed the order of V_{imp} reduces to $V_{imp} = O(\frac{1}{n})$. For a given sample size n and if M tends to infinity V_{imp} tends to zero.

4.2.2 Variance Estimation

When estimating the total variance V_{tot} we need to estimate the three components, such that:

$$\hat{V}_{tot} = \hat{V}_{sam} + \hat{V}_{imp} + \hat{V}_{res}. \tag{4.32}$$

In the following we derive an approximately unbiased estimator of the variance expression given in result 4.1.

Result 4.3

Under the assumptions of (i) equal probability sampling, (ii) uniform nonresponse within classes, (iii) n_k, the number of respondents in class B_k, and n_k, the number of units in class B_k, are large for all k, and (iv) the unweighted estimator $\hat{P}.$ is obtained under the imputation method described in section 2.2.3 selecting donors by simple random sampling with replacement,

an approximately unbiased estimator of $\text{var}_{(1)DRI}(\hat{P}.)$ is

$$\hat{\text{var}}_{(1)DRI}(\hat{P}.) = \frac{\hat{P}.(1-\hat{P}.)}{n-1} + \frac{1}{n(n-1)} \sum_k n_k \bar{z}_{.sk}(1-\bar{z}_{.sk}) \left(\frac{1}{M}(1-\hat{\pi}_k)(1-\frac{1}{n_{*k}}) + (\frac{1}{\hat{\pi}_k}-1)(1-\frac{1}{n_k}) \right)$$

$$* \left[1 - \frac{1}{n_k} \left(1 + \frac{1}{M}(1-\hat{\pi}_k)(1-\frac{1}{n_{*k}}) + (\frac{1}{\hat{\pi}_k}-1)(1-\frac{1}{n_k}) \right) \right]^{-1} \qquad (4.33)$$

where $\bar{z}_{.sk} = \frac{1}{M}\sum_{m=1}^{M} \bar{z}_{.skm}$ and $\bar{z}_{.skm} = \frac{1}{n_k}\sum_{i \in B_k} z_{.mi}$.

Proof

The proof of result 4.3 is given in the appendix A4.2.

We now use some approximations to find a simplified formula for $\hat{\text{var}}_{(1)DRI}(\hat{P}.)$.

Corollary 4.4

An approximately unbiased estimator of $\text{var}_{(2)DRI}(\hat{P}.)$, assuming that $1/n_k$ and $1/n_{*k}$ are small, is given by the following formula:

$$\hat{\text{var}}_{(2)DRI}(\hat{P}.) = \frac{\hat{P}.(1-\hat{P}.)}{n-1} + \frac{1}{n(n-1)} \sum_k n_k \bar{z}_{.sk}(1-\bar{z}_{.sk}) \left(\frac{1}{M}(1-\hat{\pi}_k) + (\frac{1}{\hat{\pi}_k}-1) \right)$$

$$= \frac{\hat{P}.(1-\hat{P}.)}{n-1} + \frac{1}{n(n-1)} \sum_k n_k \bar{z}_{.sk}(1-\bar{z}_{.sk})\,(1-\hat{\pi}_k)(\frac{1}{M}+\frac{1}{\hat{\pi}_k}) \qquad (4.34)$$

❑

Recall that the total variance is made up of three components and that the estimation of this total variance requires the (approximate) estimation of each component, i.e.

$$\hat{V}_{tot} = \hat{V}_{sam} + \hat{V}_{imp} + \hat{V}_{res} . \qquad (4.35)$$

We are now able to state simple approximately unbiased estimators of the three variance components, assuming that $1/n_k$ and $1/n_{rk}$ are small,

$$\hat{V}_{sam} = \frac{\hat{P}.(1-\hat{P}.)}{n-1} \tag{4.36}$$

$$\hat{V}_{imp} = \frac{1}{Mn(n-1)} \sum_k n_k \bar{z}_{.sk}(1-\bar{z}_{.sk})\,(1-\hat{\pi}_k) \tag{4.37}$$

$$\hat{V}_{res} = \frac{1}{n(n-1)} \sum_k n_k \bar{z}_{.sk}(1-\bar{z}_{.sk})\,(\frac{1}{\hat{\pi}_k}-1). \tag{4.38}$$

It should be noted that $\hat{V}_{sam}$ is the standard sampling variance estimator for full response treating the imputed data as real. Also, it can be shown that the formulae (4.36)-(4.38) lead to the variance formula proposed by Rao and Shao (1992) given in (4.8) in the case of hot deck imputation under uniform nonresponse, i.e. assuming one class B_k, and single imputation, M=1, where $s.^2/n$ reduces to $\hat{P}.(1-\hat{P}.)/n$ in the case of a proportion instead of a mean. To show that the two variance estimation formulae are approximately equal note that $\hat{\pi}_k = n_{rk}/n_k$, $\bar{z}_{.sk} = \hat{P}.$ if only one class such that k=1, and $\hat{V}_{res}$ in (4.8) includes both the components (4.37) and (4.38).

In the following it is of interest to consider the importance of the three variance components on the total variance $\hat{V}_{tot}$. Investigating the order in probability of the three components we see that $\hat{V}_{sam} = O_p(\frac{1}{n})$, $\hat{V}_{imp} = O_p(\frac{1}{Mn})$ and $\hat{V}_{res} = O_p(\frac{1}{n_r})$. If M is regarded as fixed the order of $\hat{V}_{imp}$ reduces to $\hat{V}_{imp} = O_p(\frac{1}{n})$. For a given sample size n the size of $\hat{V}_{imp}$ depends on the number of imputations used in the imputation process and if M tends to infinity $\hat{V}_{imp}$ tends to zero. We can see that the component $\hat{V}_{res}$ dominates the total variance if $n_r \ll n$, i.e. if the ratio $\hat{\pi}_k = (n_{rk}/n_k)$ is small. If $\hat{\pi}_k = (n_{rk}/n_k)$ is large $\hat{V}_{res}$ is of the same order as $\hat{V}_{sam}$. Note that no other assumptions about $\hat{\pi}_k = n_{rk}/n_k$ are made than $n_k = n_{rk} + n_{\bar{r}k}$ and therefore $n_{rk} \le n_k$. As an example we compare the three components for two different quarters of the LFS using the formula of $\hat{\text{var}}_{(2)DRI}(\hat{P}.)$ (table 4.1). For M=10 $\hat{V}_{imp}$ is the smallest part of the total variance when applying the formula to several quarters of the LFS. Both of the two components V_{sam} and V_{res} account for a large part of the variance, where V_{res} is larger the lower the probability of response is. Note that for quarter September-November 1999 $\hat{V}_{res}$ makes up the largest part of the total

variance, whereas for quarter March-May 2000 $\hat{V}_{res}$ is smaller than $\hat{V}_{sam}$, since the probability of response is greater for the quarter March-May 2000. The part of the variance caused by random imputation $\hat{V}_{imp}$ is with 2% of the total variance very small.

	September-November 1999	March-May 2000
$\hat{V}_{sam}$	$75.51*10^{-8}$ (40%)	$59.80*10^{-8}$ (69%)
$\hat{V}_{imp}$	$2.95*10^{-8}$ (2%)	$1.41*10^{-8}$ (2%)
$\hat{V}_{res}$	$108.11*10^{-8}$ (58%)	$25.35*10^{-8}$ (29%)

Table 4.1: Estimated variance components for two quarters of the LFS, September-November 1999 with a response rate of 25% and March-May 2000 with a response rate of 43%, for estimator $\hat{P}_1$.. The percentages indicate the percentages of the total variance $\hat{V}_{tot}$.

4.2.3 Multiple Imputation Variance Estimator

In this section we compare the performance of the multiple imputation estimator proposed by Rubin (1987) with the variance estimation formulae derived above. Using the notation in section 1.4.5 in the case of proper multiple imputation

$$\bar{G}. = \frac{1}{M}\sum_{m=1}^{M} \hat{\text{var}}_{naive}(\hat{P}_{.m}) = \frac{1}{M}\sum_{m=1}^{M} \frac{\hat{P}_{.m}(1-\hat{P}_{.m})}{n-1} \text{ for the within-imputation variance,} \tag{4.39}$$

$$\hat{\theta}. = \frac{1}{M}\sum_{m=1}^{M} \hat{\theta}.^{(m)} = \frac{1}{M}\sum_{m=1}^{M} \hat{P}_{.m} = \hat{P}. \text{ and} \tag{4.40}$$

$$\hat{B}. = \frac{1}{M-1}\sum_{m=1}^{M} (\hat{P}_{.m} - \hat{P}.)^2 \text{ for the between-imputation variance.} \tag{4.41}$$

It follows that the multiple imputation variance estimator, $\hat{T}. = \bar{G}. + (1+\frac{1}{M})\hat{B}.$, is given by

$$\hat{\mathrm{var}}_{MI}(\hat{P}_{\cdot}) = \frac{1}{M}\sum_{m=1}^{M}\frac{\hat{P}_{\cdot m}(1-\hat{P}_{\cdot m})}{n-1} + (1+\frac{1}{M})\frac{1}{M-1}\sum_{m=1}^{M}(\hat{P}_{\cdot m}-\hat{P}_{\cdot})^2 . \qquad (4.42)$$

Note that this variance estimator is based upon the assumption of proper multiple imputation, whereas the imputation method used here is based on a hot deck procedure within imputation classes repeated M times. This multiple imputation method is improper. We are therefore interested in how this variance estimation designed for proper multiple imputation performs for the improper multiple imputation method used here.

We now calculate the expectation of this variance estimator taking into account D, R and I and assuming (i) equal probability sampling, (ii) uniform nonresponse within classes, (iii) n_{rk}, the number of respondents in class B_k, and n_k, the number of units in class B_k, are large for all k, and (iv) the unweighted estimator $\hat{P}_{\cdot}$ is obtained under the imputation method described in section 2.2.3 selecting donors by simple random sampling with replacement. We have

$$E_{DRI}(\hat{\mathrm{var}}_{MI}(\hat{P}_{\cdot})) = E_{DRI}\left[\frac{1}{M}\sum_{m=1}^{M}\frac{\hat{P}_{\cdot m}(1-\hat{P}_{\cdot m})}{n-1}\right] + E_{DRI}\left[(1+\frac{1}{M})\frac{1}{M-1}\sum_{m=1}^{M}\left(\hat{P}_{\cdot m}-\hat{P}_{\cdot}\right)^2\right]$$

$$\doteq \frac{P(1-P)}{n} - \frac{b}{n-1} + (1+\frac{1}{M})\frac{1}{nN}\sum_{k} N_k \bar{z}_{Uk}(1-\bar{z}_{Uk})(1-\pi_k)$$

$$= \frac{P(1-P)}{n} - \frac{b}{n-1} + \frac{1}{MnN}\sum_{k} N_k \bar{z}_{Uk}(1-\bar{z}_{Uk})(1-\pi_k) + \frac{1}{nN}\sum_{k} N_k \bar{z}_{Uk}(1-\bar{z}_{Uk})(1-\pi_k), \qquad (4.43)$$

where $b = \frac{1}{nN}\sum_{k} N_k \bar{z}_{Uk}(1-\bar{z}_{Uk})(1-\pi_k)(1+\frac{1}{\pi_k}) = V_{imp,M=1} + V_{res}$

and $V_{imp,M=1}$ denotes the variance V_{imp} for $M=1$. The complete derivations of $E_{DRI}(\hat{\mathrm{var}}_{MI}(\hat{P}_{\cdot}))$ are given in appendix A4.3.

Comparing this result to corollary 4.2, where

$$\mathrm{var}_{(2)DRI}(\hat{P}_{\cdot}) = V_{sam} + V_{imp} + V_{res}$$

$$\doteq \frac{P(1-P)}{n} + \frac{1}{MnN}\sum_k N_k \bar{z}_{Uk}(1-\bar{z}_{Uk})(1-\pi_k) + \frac{1}{nN}\sum_k N_k \bar{z}_{Uk}(1-\bar{z}_{Uk})(\frac{1}{\pi_k}-1),$$

we can see that $\hat{\text{var}}_{MI}(\hat{P}.)$ approximates the term for V_{sam} very closely and represents the term for V_{imp} correctly. However, it does not reflect the term for V_{res} properly. This part of the variance is underestimated considerably when using the between imputation formula, since $1-\pi_k \leq \pi_k^{-1}-1$ where $0<\pi_k \leq 1$ for all k. In particular for small π_k V_{res} is underestimated. We can conclude that in the framework stated above using improper multiple imputation and taking into account imputation, response and sampling variability the variance estimation using the multiple imputation formula by Rubin is biased. It underestimates the true variance since it underestimates V_{res}. This is because the draws from the hot deck do not represent the full uncertainty in estimating the data for purposes of multiple variance estimation (see also Rao, 1996; Fay, 1996).

We now consider correcting for the bias of the multiple imputation variance estimator. It follows from (8.5) in the appendix that V_{sam} can be estimated by the within imputation variance $M^{-1}\sum_{m=1}^{M}\hat{P}_{.m}(1-\hat{P}_{.m})/(n-1)$, since $b/(n-1)=O(1/n^2)$, whereas $V_{sam}=O(1/n)$. From (8.6) in the appendix it follows that V_{imp} can be approximately estimated by the between-imputation variance $\hat{B}$ times a factor depending on the number of multiple imputations M, i.e. it can be estimated by

$$\frac{1}{M(M-1)}\sum_{m=1}^{M}(\hat{P}_{.m}-\hat{P}.)^2 . \qquad (4.44)$$

The component V_{res} is not estimated correctly by $\hat{\text{var}}_{MI}(\hat{P}.)$ such that this component is estimated by $\hat{V}_{res}$ given in (4.38). Since the multiple imputation formula by Rubin does not capture the full response variability but reflects the sampling and the imputation variance correctly, we can use the following modification to Rubin's formula to obtain an approximately unbiased formula for the framework used here.

Result 4.5

Under the assumptions of (i) equal probability sampling, (ii) uniform nonresponse within classes, (iii) n_{rk}, the number of respondents in class B_k, and n_k, the number of units in class B_k, are large

for all k, and (iv) the unweighted estimator $\hat{P}_{\cdot}$ is obtained under the imputation method described in section 2.2.3 selecting donors by simple random sampling with replacement, a simplified approximately unbiased estimator for $\mathrm{var}_{(2)DRI}(\hat{P}_{\cdot})$ is given by the following formula taking into account a modification of Rubin's rule:

$$\hat{\mathrm{var}}'_{MI}(\hat{P}_{\cdot}) = \frac{1}{M}\sum_{m=1}^{M}\frac{\hat{P}_{\cdot m}(1-\hat{P}_{\cdot m})}{n-1} + \frac{1}{M(M-1)}\sum_{m=1}^{M}(\hat{P}_{\cdot m}-\hat{P}_{\cdot})^2$$

$$+\frac{1}{n(n-1)}\sum_{k} n_k \bar{z}_{\cdot sk}(1-\bar{z}_{\cdot sk})\left(\frac{1}{\hat{\pi}_k}-1\right)$$

$$= \bar{G}_{\cdot} + \frac{1}{M}\hat{B}_{\cdot} + \hat{V}_{res}, \qquad (4.45)$$

where $\bar{z}_{\cdot sk} = \frac{1}{M}\sum_{m=1}^{M}\bar{z}_{\cdot skm} = \frac{1}{M}\sum_{m=1}^{M}\frac{1}{n_k}\sum_{i\in B_k} z_{\cdot mi}$.

❑

However, since this modification of the multiple imputation formula include calculations for each imputation class the modification does not gain much for practical applications. We therefore still recommend $\hat{\mathrm{var}}_{(2)DRI}(\hat{P}_{\cdot})$ for practical use.

4.3 Variance and Variance Estimation Allowing For Fixed Survey Weights

In the following section we derive a valid variance formula for the estimator $\hat{P}_{\cdot}$ and give an approximately unbiased estimator of this variance taking into account the complex weighting scheme of the LFS allowing for fixed weights as in section 3.2. The required framework and necessary definitions for the following results are given in sections 3.1 and 3.2. Note that all formulae derived here reduce to the simpler case described in section 4.2 when the weights are constant.

In addition, we define for the weights:

$$\overline{w}_{Uk} = N_k^{-1} \sum_{i \in B_{Uk}} w_i , \qquad \overline{w_s^2} = \frac{1}{\sum_{i \in s} w_i} \sum_{i \in s} w_i^2 ,$$

$$\overline{w_{sk}^2} = \frac{1}{\sum_{i \in B_k} w_i} \sum_{i \in B_k} w_i^2 , \quad \overline{w_{rk}^2} = \frac{1}{\sum_{i \in B_{k \cap r}} w_i} \sum_{i \in B_{k \cap r}} w_i^2 . \tag{4.46}$$

4.3.1 Variance of the Estimator $\hat{P}_.$

Taking into account the survey weights $\hat{P}_.$, as defined in section 2.2.5, is a non-linear estimator. We define

$$\hat{P}_{.m} = \frac{\sum_{i \in s} w_i z_{.mi}}{\sum_{i \in s} w_i} = \frac{u_m}{v} = g(u_m, v) . \tag{4.47}$$

When deriving the variance of a non-linear estimator we first need to approximate the estimator by using Taylor series methods (delta-method) (Bishop, Fienberg, and Holland, 1978). The estimator based on the mth imputation $\hat{P}_{.m}$ can be approximated by:

$$\hat{P}_{.m} = g(u_m, v) \doteq g(U,V) + (u_m - U)\frac{1}{V} + (v - V)\frac{-U}{V^2} , \tag{4.48}$$

where $E_{DRI}(u_m) = U$ and $E_{DRI}(v) = E_D(v) = V$, $u_m = \sum_{i \in s} w_i z_{.mi}$, $U = \sum_{i \in U} z_i$, $v = \sum_{i \in s} w_i$ and $V = N$ (see also proof of result 3.2).

We can therefore approximate the variance by

$$\mathrm{var}_{DRI}(\hat{P}_.) = \mathrm{var}_{DRI}\left(\frac{1}{M}\sum_{m=1}^{M} g(u_m, v)\right) \doteq \mathrm{var}_{DRI}\left(\frac{1}{M}\sum_{m=1}^{M} g(U,V) + (u_m - U)\frac{1}{V} + (v - V)\frac{-U}{V^2}\right)$$

$$= \frac{1}{V^2}\mathrm{var}_{DRI}\left(\frac{1}{M}\sum_{m=1}^{M}(u - v\frac{U}{V})\right) = \frac{1}{N^2}\mathrm{var}_{DRI}\left(\frac{1}{M}\sum_{m=1}^{M}\sum_{i \in s} w_i (z_{.mi} - P)\right) . \tag{4.49}$$

Result 4.6

Under the assumptions of (i) equal probability sampling, (ii) uniform nonresponse within classes, (iii) the number of respondents and the number of units in class B_k are large for all k, and (iv) the weighted estimator $\hat{P}.$ is obtained under the imputation method described in section 2.2.3 selecting donors with replacement,

the variance formula that takes into account imputation, sampling and nonresponse variability and the weighting scheme of the LFS, ignoring the finite population correction, is approximately given by

$$\mathrm{var}_{DRI}(\hat{P}.) \doteq \frac{1}{N^2}\left\{\sum_{i\in U} w_i (z_i - P)^2 + \frac{1}{M}\sum_k N_k \bar{z}_{Uk}(1-\bar{z}_{Uk})(1-\pi_k)\overline{w_{Uk}} \right.$$

$$\left. + \sum_k \sum_{i\in B_{Uk}} (\frac{1}{\pi_k} - 1) w_i (z_i - \bar{z}_{Uk})^2 \right\} \tag{4.50}$$

Proof

The proof is given in appendix A4.4.

As shown in this proof the total variance consists of three components:

$$\mathrm{var}_{DRI}(\hat{P}.) = V_{tot} = V_{sam} + V_{imp} + V_{res}, \tag{4.51}$$

where

$$V_{sam} \doteq \frac{1}{N^2}\sum_{i\in U} w_i (z_i - P)^2 \tag{4.52}$$

$$V_{imp} \doteq \frac{1}{N^2}\frac{1}{M}\sum_k N_k \bar{z}_{Uk}(1-\bar{z}_{Uk})(1-\pi_k)\overline{w_{Uk}} \tag{4.53}$$

$$V_{res} \doteq \frac{1}{N^2}\sum_k \sum_{i\in B_{Uk}} (\frac{1}{\pi_k} - 1) w_i (z_i - \bar{z}_{Uk})^2 . \tag{4.54}$$

Result 4.7:

The approximations of the three components V_{sam}, V_{imp} and V_{res} given in (4.52)-(4.54) taking into account fixed survey weights reduce to the approximations of V_{sam}, V_{imp} and V_{res} given in (4.29)-(4.31), not taking into account the survey weights, when these weights are equal.

Proof

Since we assume $w_i = 1/v_i$ and $v_i = E_D(I(i \in s)) = n/N$ we have $w_i = N/n$. It follows

$$V_{sam} \doteq \frac{1}{N^2}\sum_{i\in U} w_i (z_i - P)^2 = \frac{1}{Nn}\sum_{i\in U}(z_i - P)^2 = \frac{1}{Nn} NP(1-P) = \frac{P(1-P)}{n} \quad \text{as in (4.29)}$$

$$V_{imp} \doteq \frac{1}{N^2}\frac{1}{M}\sum_k (N_k \bar{z}_{Uk}(1-\bar{z}_{Uk})(1-\pi_k)\frac{1}{N_k}\sum_{i\in B_{Uk}} w_i)$$

$$= \frac{1}{MnN}\sum_k N_k \bar{z}_{Uk}(1-\bar{z}_{Uk})(1-\pi_k) \quad \text{as in (4.30)}$$

$$V_{res} \doteq \frac{1}{N^2}\sum_k \sum_{i\in B_{Uk}} (\frac{1}{\pi_k} - 1) w_i (z_i - \bar{z}_{Uk})^2 = \frac{1}{nN}\sum_k N_k \bar{z}_{Uk}(1-\bar{z}_{Uk})(\frac{1}{\pi_k} - 1) \quad \text{as in (4.31)}$$

❑

4.3.2 Variance Estimation

An approximately unbiased estimator of the variance allowing for fixed survey weights is derived.

Result 4.8

An approximately unbiased estimator of $\text{var}_{DRI}(\hat{P}_.)$ is given by

$$\hat{\text{var}}_{DRI}(\hat{P}_.) = \frac{1}{(\sum_{i\in s} w_i)^2} * \left\{ \frac{1}{M}\sum_{m=1}^{M}\sum_k \sum_{i\in B_k} \frac{1}{\hat{\pi}_k} w_i^2 \left((z_{.mi} - \hat{P}_.)^2 + (z_{.mi} - \bar{z}_{.sk})^2 (\frac{1}{\hat{\pi}_k} - 1) \right) \right.$$

$$\left. + \frac{1}{M}\sum_k \bar{z}_{.sk}(1-\bar{z}_{.sk})(1-\hat{\pi}_k)\sum_{i\in B_k} w_i^2 \right.$$

$$-\sum_k \sum_{i\in B_k} \left(\frac{1}{\hat{\pi}_k} - 1\right) w_i^2 \left[(\bar{z}_{\cdot sk} - \hat{P}_{\cdot})^2 + \bar{z}_{\cdot sk}(1-\bar{z}_{\cdot sk}) \frac{1}{\hat{\pi}_k} \right] \Bigg\} \tag{4.55}$$

Proof

The proof of result 4.8 is given in A4.5.

Similarly to result 4.7 it can be shown that $\hat{\text{var}}_{DRI}(\hat{P}_{\cdot})$ not taking into account the survey weights, when these weights are equal, reduces to $\hat{\text{var}}_{(2)DRI}(\hat{P}_{\cdot})$ given in corollary 4.4.

4.4 Simulation Study for Variance Estimation without Weighting Adjustment

To investigate the performance of the variance estimation formulae derived in section 4.2 a simulation study is carried out under the same conditions as described in section 3.3. For simplicity the survey weights w_i are not taken into account. In the following we will discuss the results of the variance estimation for point estimator $\hat{P}_{1\cdot}$. The following variance formulae are investigated, where the superscript a refers to each generated sample $s^{(a)}$: the newly derived approximately unbiased variance formulae $\hat{\text{var}}^{(a)}_{(1)DRI}(\hat{P}_{1\cdot}^{(a)})$ and $\hat{\text{var}}^{(a)}_{(2)DRI}(\hat{P}_{1\cdot}^{(a)})$, the multiple imputation variance formula $\hat{\text{var}}^{(a)}_{MI}(\hat{P}_{1\cdot}^{(a)})$ and the approximately unbiased modification of this formula $\hat{\text{var}}'^{(a)}_{MI}(\hat{P}_{1\cdot}^{(a)})$. We also analyse the naive variance that applies for full response

$$\text{var}^{(a)}_{naive}(\hat{P}_{1\cdot}^{(a)}) = \frac{\hat{P}_{1\cdot}^{(a)}(1-\hat{P}_{1\cdot}^{(a)})}{n}. \tag{4.56}$$

To generate a nonresponse mechanism following the MAR assumption only Model A3 was used. In the following analysis only point estimator $\hat{P}_{1\cdot}$ is considered. The sampling fraction is assumed to be negligible. The performances of the different formulae are assessed by the following criteria (see also Lee, Rancourt and Särndal, 1994 and 2000).

i.) Estimated Bias of the Variance Estimator

$$\hat{Bias}(\hat{V}(\hat{P}.)) = E_s(\hat{V}) - V = \frac{1}{A}\sum_{a=1}^{A}\hat{V}^{(a)} - V\,, \tag{4.57}$$

where $V = \frac{1}{A-1}\sum_{a=1}^{A}(\hat{P}.^{(a)} - \bar{P})^2$, $\bar{P} = \frac{1}{A}\sum_{a=1}^{A}\hat{P}.^{(a)}$, is the simulation variance of the imputed estimator $\hat{P}.$ and $\hat{V}^{(a)}$ is the variance estimate in simulation a.

ii.) Estimated Relative Bias of the Variance Estimator

$$\hat{RB}(\hat{V}(\hat{P}.)) = 100 * \frac{E_S(\hat{V}) - V}{V}\,. \tag{4.58}$$

iii.) Estimated Root Mean Square Error

$$\hat{RMSE}(\hat{V}(\hat{P}.)) = \sqrt{E_S((\hat{V}^{(a)} - V)^2)} = \sqrt{\frac{1}{A}\sum_{a=1}^{A}(\hat{V}^{(a)} - V)^2}\,. \tag{4.59}$$

iv.) Estimated Relative Root Mean Square Error

$$\hat{RRMSE}(\hat{V}(\hat{P}.)) = \frac{\hat{RMSE}(\hat{V}(\hat{P}.))}{V} * 100\,. \tag{4.60}$$

v.) Simulation Variance of the Variance Estimator

$$V(\hat{V}(\hat{P}.)) = \frac{1}{A-1}\sum_{a=1}^{A}(\hat{V}^{(a)} - \bar{V})^2\,, \tag{4.61}$$

where $\bar{V} = \frac{1}{A}\sum_{a=1}^{A}\hat{V}^{(a)}$.

vi.) Relative Simulation Variance of the Variance Estimator

$$RV(\hat{V}(\hat{P}.)) = \frac{V(\hat{V}(\hat{P}.))}{V} * 100\,. \tag{4.62}$$

vii.) Average Length of the 95%-Confidence Interval

Let $b_{V(\hat{P}.)}$ be the lower bound and $c_{V(\hat{P}.)}$ be the upper bound of the 95% confidence interval based on using $V(\hat{P}.)$ then

$$AL = \frac{1}{A}\sum_{a=1}^{A} | b^{(a)}_{V(\hat{P}_\cdot)} - c^{(a)}_{V(\hat{P}_\cdot)} | \tag{4.63}$$

denotes the average length of the 95% confidence interval.

viii) Variance of the Length of the 95%-Confidence Interval

The variance of the length of the 95% confidence interval is defined as

$$\frac{1}{A}\sum_{a=1}^{A} | b^{(a)}_{V(\hat{P}_\cdot)} - c^{(a)}_{V(\hat{P}_\cdot)} | - AL)^2 \tag{4.64}$$

ix.) Coverage Rate

The 95% confidence interval $\hat{P}_\cdot^{(a)} \pm 1.96\sqrt{\hat{V}^{(a)}}$ is calculated for each simulation a. The coverage rate (COVR) is defined as the ratio of the number of times that the confidence interval covers the true value P divided by the total number of iterations A. Note that the confidence intervals using $\hat{\text{var}}_{MI}(\hat{P}_\cdot)$ are derived as discussed in chapter 1, section 1.4.5.

4.4.1 Results for Variance Estimators under Different Response Mechanisms

The results for the variance formulae under three different response mechanisms are given in tables 4.2 and 4.3 for hot deck imputation within classes sampling donors by SRS with replacement. We can see that for imputation with replacement the newly derived variance estimators $\hat{\text{var}}_{(1)DRI}(\hat{P}_1\cdot)$ and $\hat{\text{var}}_{(2)DRI}(\hat{P}_1\cdot)$ perform very well under uniform and uniform within classes response mechanism, with an (estimated) relative bias of under 2% (table 4.2). They show a coverage rate very close to 95% for the 95%-confidence interval (table 4.3). Under MAR nonresponse based on model A3, the relative bias is around 3% and the coverage rates are close to 95%. We can therefore conclude that the newly derived variance estimators perform well under all three nonresponse mechanisms and are reasonably robust to departures of the assumption of uniform response within imputation classes.

The variance formula for multiple imputation performs badly with a relative bias between 10% and 30% depending on the response mechanism. The coverage rate is around 90% under uniform

and uniform within classes nonresponse and therefore lower than the necessary 95%. Only for MAR nonresponse the coverage rate is close to 95%, which might be related to the fact that under the MAR nonresponse considered here the response rate at the lower end of the distribution of Y is higher than under uniform nonresponse. Using the modified formula $\hat{\text{var}}'_{MI}(\hat{P}_1.)$ instead of $\hat{\text{var}}_{MI}(\hat{P}_1.)$ shows very good results for all three response mechanisms, with a relative bias between 0.7% and 3% and coverage rates close to 95%.

	Variance Formula	Bias	Rel. Bias	RMSE	Rel. RMSE
uniform response	$\hat{\text{var}}_{(1)DRI}(\hat{P}_1.)$	$-0.46*10^{-7}$	-0.64 %	$3.12*10^{-7}$	4.3 %
	$\hat{\text{var}}_{(2)DRI}(\hat{P}_1.)$	$-0.49*10^{-7}$	-0.69 %	$3.13*10^{-7}$	4.4 %
	$\hat{\text{var}}_{MI}(\hat{P}_1.)$	$-19.53*10^{-7}$	-27.47%	$20.91*10^{-7}$	29.4 %
	$\hat{\text{var}}'_{MI}(\hat{P}_1.)$	$-0.48*10^{-7}$	-0.68 %	$3.21*10^{-7}$	4.5 %
	$\hat{\text{var}}_{naive}(\hat{P}_1.)$	$-35.41*10^{-7}$	-49.90 %	$35.41*10^{-7}$	49.8 %
uniform within classes	$\hat{\text{var}}_{(1)DRI}(\hat{P}_1.)$	$-1.19*10^{-7}$	-1.69 %	$3.07*10^{-7}$	4.3 %
	$\hat{\text{var}}_{(2)DRI}(\hat{P}_1.)$	$-1.23*10^{-7}$	-1.75 %	$3.09*10^{-7}$	4.4 %
	$\hat{\text{var}}_{MI}(\hat{P}_1.)$	$-20.36*10^{-7}$	-28.94 %	$21.51*10^{-7}$	30.6 %
	$\hat{\text{var}}'_{MI}(\hat{P}_1.)$	$-1.25*10^{-7}$	-1.79 %	$3.16*10^{-7}$	4.5 %
	$\hat{\text{var}}_{naive}(\hat{P}_1.)$	$-35.01*10^{-7}$	-49.90 %	$35.02*10^{-7}$	49.8 %
MAR (Model A3)	$\hat{\text{var}}_{(1)DRI}(\hat{P}_1.)$	$-1.34*10^{-7}$	-2.85 %	$2.16*10^{-7}$	4.6 %
	$\hat{\text{var}}_{(2)DRI}(\hat{P}_1.)$	$-1.34*10^{-7}$	-2.86 %	$2.16*10^{-7}$	4.6 %
	$\hat{\text{var}}_{MI}(\hat{P}_1.)$	$-3.84*10^{-7}$	-8.16 %	$5.43*10^{-7}$	11.5 %
	$\hat{\text{var}}'_{MI}(\hat{P}_1.)$	$-1.32*10^{-7}$	-2.82 %	$2.17*10^{-7}$	4.6 %
	$\hat{\text{var}}_{naive}(\hat{P}_1.)$	$-11.32*10^{-7}$	-24.16 %	$11.43*10^{-7}$	24.2 %

Table 4.2: Simulation results for variance estimators for hot deck imputation with replacement.

	Variance Formula	Simulation Variance of Variance Estimator	Rel. Simulation Variance of Variance Estimator	Average Length of CI	Variance of Length of CI	Coverage Rate
uniform response	$\hat{var}_{(1)DRI}(\hat{P}_1.)$	$9.82*10^{-14}$	$1.31*10^{-6}$ %	$10.0*10^{-3}$	$0.53*10^{-7}$	94.7 %
	$\hat{var}_{(2)DRI}(\hat{P}_1.)$	$9.83*10^{-14}$	$1.31*10^{-6}$ %	$10.0*10^{-3}$	$0.53*10^{-7}$	94.7 %
	$\hat{var}_{MI}(\hat{P}_1.)$	$61.82*10^{-14}$	$8.25*10^{-6}$ %	$8.9*10^{-3}$	$5.59*10^{-7}$	89.4 %
	$\hat{var}'_{MI}(\hat{P}_1.)$	$1.02*10^{-14}$	$1.36*10^{-6}$ %	$10.0*10^{-3}$	$0.56*10^{-7}$	94.8 %
	$\hat{var}_{naive}(\hat{P}_1.)$	$2.62*10^{-14}$	$0.35*10^{-6}$ %	$7.4*10^{-3}$	$0.28*10^{-7}$	83.0 %
uniform within classes	$\hat{var}_{(1)DRI}(\hat{P}_1.)$	$2.65*10^{-14}$	$0.61*10^{-6}$ %	$8.3*10^{-3}$	$0.22*10^{-7}$	94.6 %
	$\hat{var}_{(2)DRI}(\hat{P}_1.)$	$2.65*10^{-14}$	$0.61*10^{-6}$ %	$8.3*10^{-3}$	$0.22*10^{-7}$	94.6 %
	$\hat{var}_{MI}(\hat{P}_1.)$	$13.38*10^{-14}$	$3.05*10^{-6}$ %	$8.1*10^{-3}$	$1.33*10^{-7}$	89.7 %
	$\hat{var}'_{MI}(\hat{P}_1.)$	$2.74*10^{-14}$	$0.62*10^{-6}$ %	$8.3*10^{-3}$	$0.23*10^{-7}$	94.6 %
	$\hat{var}_{naive}(\hat{P}_1.)$	$1.52*10^{-14}$	$0.35*10^{-6}$ %	$7.4*10^{-3}$	$0.16*10^{-7}$	83.1 %
MAR (Model A3)	$\hat{var}_{(1)DRI}(\hat{P}_1.)$	$2.69*10^{-14}$	$0.59*10^{-6}$ %	$8.3*10^{-3}$	$0.21*10^{-7}$	94.8 %
	$\hat{var}_{(2)DRI}(\hat{P}_1.)$	$2.69*10^{-14}$	$0.59*10^{-6}$ %	$8.3*10^{-3}$	$0.22*10^{-7}$	94.8 %
	$\hat{var}_{MI}(\hat{P}_1.)$	$15.81*10^{-14}$	$3.50*10^{-6}$ %	$8.1*10^{-3}$	$1.58*10^{-7}$	94.6 %
	$\hat{var}'_{MI}(\hat{P}_1.)$	$2.85*10^{-14}$	$0.65*10^{-6}$ %	$8.3*10^{-3}$	$0.24*10^{-7}$	94.9 %
	$\hat{var}_{naive}(\hat{P}_1.)$	$1.59*10^{-14}$	$0.35*10^{-6}$ %	$7.4*10^{-3}$	$0.17*10^{-7}$	92.0 %

Table 4.3: Simulation results for variance estimators for hot deck imputation with replacement.

The relative simulation variance of the variance estimator is largest for $\hat{var}_{MI}(\hat{P}_1.)$ under uniform and uniform within classes nonresponse, whereas the formulae $\hat{var}_{(1)DRI}(\hat{P}_1.)$, $\hat{var}_{(2)DRI}(\hat{P}_1.)$ and $\hat{var}'_{MI}(\hat{P}_1.)$ perform better showing a smaller simulation variance.

The variance of the length of the confidence interval is also much smaller for $\hat{var}_{(1)DRI}(\hat{P}_1.)$, $\hat{var}_{(2)DRI}(\hat{P}_1.)$ and $\hat{var}'_{MI}(\hat{P}_1.)$ than for $\hat{var}_{MI}(\hat{P}_1.)$. As expected the naive variance estimator performs the worst, underestimating the true variance with a relative bias between 25% and 50% and a coverage rate between 83% and 92%, depending on the response mechanism. We can also see that the approximations made to $\hat{var}_{(1)DRI}(\hat{P}_1.)$, using $\hat{var}_{(2)DRI}(\hat{P}_1.)$ instead of $\hat{var}_{(1)DRI}(\hat{P}_1.)$, work well. Note that computing $\hat{var}_{(1)DRI}(\hat{P}_1.)$ and $\hat{var}_{(2)DRI}(\hat{P}_1.)$ for several quarters of the Labour Force Survey showed that $\hat{var}_{(2)DRI}(\hat{P}_1.)$ is a very good approximation to $\hat{var}_{(1)DRI}(\hat{P}_1.)$. The results are presented in table 4.4. We therefore conclude that the simplified formula $\hat{var}_{(2)DRI}(\hat{P}_1.)$ instead of $\hat{var}_{(1)DRI}(\hat{P}_1.)$, not taking into account the survey weights, can be used in practice, under the conditions discussed here.

LFS quarter	$\hat{var}_{(1)DRI}(\hat{P}_1.)$	$\hat{var}_{(2)DRI}(\hat{P}_1.)$
June-August 1999	$1.8428*10^{-6}$	$1.8486*10^{-6}$
Sept-Nov. 1999	$1.9255*10^{-6}$	$1.9380*10^{-6}$
Dec. 1999-Jan. 2000	$1.4102*10^{-6}$	$1.4391*10^{-6}$

Table 4.4: Variance estimates based on $\hat{var}_{(1)DRI}(\hat{P}_1.)$ and $\hat{var}_{(2)DRI}(\hat{P}_1.)$ for several quarters of the LFS.

Tables 4.5 and 4.6 present the results for hot deck imputation within classes sampling donors by SRS without replacement. Under selection of donors without replacement in the imputation process the newly derived variance formulae, which are developed under the assumption of imputation with replacement, still perform well with a relative bias of under 2% (table 4.5). They now give a more conservative result, overestimating the true variance slightly. This is expected since selecting donors without replacement instead of with replacement decreases the variance due to imputation. This has the consequence that the component for $\hat{V}_{imp}$ in the variance formula overestimates the true imputation variance V_{imp} slightly. The results for the formula for multiple

imputation according to Rubin are worse for imputing without replacement under all three response mechanisms. The modification $\hat{var}'_{MI}(\hat{P}_1.)$ performs very well with a coverage rate close to 95% and a relative bias of under 1% for all three response mechanisms. There is an indication that in the case of imputation without replacement this formula performs better than $\hat{var}_{(1)DRI}(\hat{P}_1.)$ and $\hat{var}_{(2)DRI}(\hat{P}_1.)$, which is assumed to be related to a better estimation of V_{imp} using $\hat{var}'_{MI}(\hat{P}_1.)$. As before the naive variance formula severely underestimates the true variance with a relative bias between -21% and -50% and a coverage rate between 85% and 90%.

	Variance formula	Bias	Rel. Bias	RMSE	Rel. RMSE
uniform response	$\hat{var}_{(1)DRI}(\hat{P}_1.)$	$1.23*10^{-7}$	1.77 %	$3.32*10^{-7}$	4.7 %
	$\hat{var}_{(2)DRI}(\hat{P}_1.)$	$1.20*10^{-7}$	1.72 %	$3.30*10^{-7}$	4.7 %
	$\hat{var}_{MI}(\hat{P}_1.)$	$-31.21*10^{-7}$	-44.90 %	$31.11*10^{-7}$	44.9 %
	$\hat{var}'_{MI}(\hat{P}_1.)$	$-0.01*10^{-7}$	-0.01 %	$3.03*10^{-7}$	4.3 %
	$\hat{var}_{naive}(\hat{P}_1.)$	$-33.83*10^{-7}$	-48.61 %	$3.38*10^{-6}$	48.6 %
uniform within classes	$\hat{var}_{(1)DRI}(\hat{P}_1.)$	$1.21*10^{-7}$	1.78 %	$3.02*10^{-7}$	4.4 %
	$\hat{var}_{(2)DRI}(\hat{P}_1.)$	$1.17*10^{-7}$	1.72 %	$3.01*10^{-7}$	4.4 %
	$\hat{var}_{MI}(\hat{P}_1.)$	$-31.3*10^{-7}$	-46.01 %	$3.13*10^{-6}$	46.1 %
	$\hat{var}'_{MI}(\hat{P}_1.)$	$-0.06*10^{-7}$	-0.08 %	$2.72*10^{-7}$	3.3 %
	$\hat{var}_{naive}(\hat{P}_1.)$	$-32.72*10^{-7}$	-48.10 %	$3.28*10^{-6}$	48.2 %
MAR (Model A3)	$\hat{var}_{(1)DRI}(\hat{P}_1.)$	$0.21*10^{-7}$	0.45 %	$1.65*10^{-7}$	3.3 %
	$\hat{var}_{(2)DRI}(\hat{P}_1.)$	$0.21*10^{-7}$	0.45 %	$1.65*10^{-7}$	3.3 %
	$\hat{var}_{MI}(\hat{P}_1.)$	$-5.47*10^{-7}$	-12.02 %	$5.96*10^{-7}$	12.2 %
	$\hat{var}'_{MI}(\hat{P}_1.)$	$-0.07*10^{-7}$	-0.14 %	$3.76*10^{-7}$	7.7 %
	$\hat{var}_{naive}(\hat{P}_1.)$	$-9.81*10^{-7}$	-21.52 %	$1.31*10^{-6}$	26.8 %

Table 4.5: Simulation results for variance estimators for hot deck imputation without replacement.

	Variance formula	Simulation Variance of Variance Estimator	Rel. Simulation Variance of Variance Estimator	Average Length of CI	Variance of Length of CI	Coverage Rate
uniform response	$\hat{var}_{(1)DRI}(\hat{P}_1.)$	$10.02*10^{-14}$	$1.34*10^{-6}$ %	$10.0*10^{-3}$	$5.43*10^{-8}$	95.0 %
	$\hat{var}_{(2)DRI}(\hat{P}_1.)$	$10.01*10^{-14}$	$1.34*10^{-6}$ %	$10.0*10^{-3}$	$5.43*10^{-8}$	95.0 %
	$\hat{var}_{MI}(\hat{P}_1.)$	$4.14*10^{-14}$	$0.56*10^{-6}$ %	$7.6*10^{-3}$	$4.27*10^{-8}$	86.2 %
	$\hat{var}'_{MI}(\hat{P}_1.)$	$9.70*10^{-14}$	$1.30*10^{-6}$ %	$9.3*10^{-3}$	$5.36*10^{-8}$	94.9 %
	$\hat{var}_{naive}(\hat{P}_1.)$	$2.59*10^{-14}$	$0.35*10^{-6}$ %	$7.4*10^{-3}$	$2.78*10^{-8}$	84.6 %
uniform within classes	$\hat{var}_{(1)DRI}(\hat{P}_1.)$	$2.72*10^{-14}$	$0.59*10^{-6}$ %	$8.3*10^{-3}$	$2.28*10^{-8}$	95.2 %
	$\hat{var}_{(2)DRI}(\hat{P}_1.)$	$2.72*10^{-14}$	$0.59*10^{-6}$ %	$8.3*10^{-3}$	$2.28*10^{-8}$	95.2 %
	$\hat{var}_{MI}(\hat{P}_1.)$	$5.42*10^{-14}$	$1.19*10^{-6}$ %	$7.8*10^{-3}$	$5.56*10^{-8}$	85.1 %
	$\hat{var}'_{MI}(\hat{P}_1.)$	$2.66*10^{-14}$	$0.58*10^{-6}$ %	$8.3*10^{-3}$	$2.25*10^{-8}$	94.9 %
	$\hat{var}_{naive}(\hat{P}_1.)$	$1.59*10^{-14}$	$0.35*10^{-6}$ %	$7.4*10^{-3}$	$1.71*10^{-8}$	84.6 %
MAR (Model A3)	$\hat{var}_{(1)DRI}(\hat{P}_1.)$	$2.52*10^{-14}$	$0.59*10^{-6}$ %	$8.3*10^{-3}$	$2.13*10^{-8}$	94.5 %
	$\hat{var}_{(2)DRI}(\hat{P}_1.)$	$2.52*10^{-14}$	$0.59*10^{-6}$ %	$8.3*10^{-3}$	$2.13*10^{-8}$	94.5 %
	$\hat{var}_{MI}(\hat{P}_1.)$	$5.97*10^{-14}$	$1.41*10^{-6}$ %	$7.8*10^{-3}$	$6.15*10^{-8}$	93.0 %
	$\hat{var}'_{MI}(\hat{P}_1.)$	$2.51*10^{-14}$	$0.59*10^{-6}$ %	$8.3*10^{-3}$	$2.13*10^{-8}$	94.4 %
	$\hat{var}_{naive}(\hat{P}_1.)$	$1.47*10^{-14}$	$0.35*10^{-6}$ %	$7.4*10^{-3}$	$1.59*10^{-8}$	90.9%

Table 4.6: Simulation results for variance estimators for hot deck imputation without replacement.

From table 4.6 we can see that the variance of the length of the confidence interval is smaller for $\hat{var}_{(1)DRI}(\hat{P}_1.)$, $\hat{var}_{(2)DRI}(\hat{P}_1.)$ and $\hat{var}'_{MI}(\hat{P}_1.)$ than for $\hat{var}_{MI}(\hat{P}_1.)$ under uniform within classes and

MAR nonresponse. The same applies to the simulation variance of the variance estimator. The average length of the confidence interval is smaller for $\hat{\text{var}}_{MI}(\hat{P}_1.)$ than for $\hat{\text{var}}_{(1)DRI}(\hat{P}_1.)$, $\hat{\text{var}}_{(2)DRI}(\hat{P}_1.)$ and $\hat{\text{var}}'_{MI}(\hat{P}_1.)$.

The simulation study shows that the proposed variance estimators $\hat{\text{var}}_{(1)DRI}(\hat{P}_1.)$ and $\hat{\text{var}}_{(2)DRI}(\hat{P}_1.)$ have small relative bias and the associated confidence intervals have good coverage under a range of conditions. It is shown that even under violations of certain assumptions the proposed variance estimators perform well. The adjusted MI variance estimator, $\hat{\text{var}}'_{MI}(\hat{P}_1.)$, also shows good results in the simulation study. For computational simplicity, however, $\hat{\text{var}}_{(2)DRI}(\hat{P}_1.)$ is recommended for practical use. The multiple imputation variance estimator $\hat{\text{var}}_{MI}(\hat{P}_1.)$ designed for proper multiple imputations does not perform well under the improper multiple imputation considered here. As expected, the naive variance formula is found not adequate for estimating the variance of a point estimator in the presence of imputation leading to severe underestimation.

4.5 The Approximate Bayesian Bootstrap

4.5.1 The Approximate Bayesian Bootstrap

In the previous section we have seen that the multiple imputation variance estimator $\hat{\text{var}}_{MI}(\hat{P}.)$ underestimates the true variance. The underestimation occurs because the imputation method considered here assumes that the parameters of the predictive distribution are known. The method therefore does not account for uncertainty in the estimates of parameters used in the linear regression equation. Several approaches exist to refine a single value imputation method into proper multiple imputation, such that the variance of an estimator can by calculated easily. The method of the approximate Bayesian Bootstrap (ABB) (Rubin and Schenker, 1986; Efron, 1994; Little and Rubin, 2002) allows for uncertainty in estimating these parameters and produces proper multiple imputations. In this section the ABB is briefly reviewed. A simulation study is carried out to evaluate the performance of the imputation method under the ABB including variance estimation based on the multiple imputation variance formula.

The ABB in the case of hot deck imputation within classes is described in Rubin (1987), Fay (1996), Rao and Shao (1992) and Kim (2002). Let the original sample contain K imputation classes or cells for example defined by the values of fully observed categorical variables. For each imputation set the donors within each imputation class are sampled (bootstrapped) with replacement of the same size as respondents are available in each class. The set of bootstrapped respondents in class B_k for the mth imputation is denoted $r_k^{(m)}$, $m=1,...,M$. For each nonrespondent in class B_k one donor is selected with replacement from the set of respondents $r_k^{(m)}$ for that class at random. This method generates M proper multiple imputations.

The case of predictive mean matching imputation and its modification to a proper multiple imputation method are described in Heitjan and Little (1991). The basic imputation method that is used is random nearest neighbour imputation, where based on the predicted values the five nearest respondents to each nonrespondent are selected and one donor value is selected for imputation out of the five values at random. This procedure is repeated to obtain M multiple imputations. However, since the method does not account for uncertainty in the parameter estimates it is an improper multiple imputation method. Heitjan and Little (1991) propose several adjustments to the imputation method to account for the additional source of variability. The basic idea is to use different sets of regression parameter estimates and each set is used to generate a single imputation. One possibility is to bootstrap the respondent cases in sample s and refit the regression model for each bootstrap sample of the respondent cases. Heitjan and Little (1991) carried out a simulation study comparing the effects of proper and improper multiple imputations. The effect of going to the proper method based on the ABB was "generally not great" and the proper method was found to be too conservative. Rao (1996) also notes limitations of the ABB. For example methods incorporating clustering, stratification and weighting compensating for unequal probabilities of selection are not currently available under the ABB. Kim (2002) investigates the finite sample properties of the variance estimator if the ABB has been used and proposes an adjustment to the formula reducing the bias of the variance estimator for finite sample sizes.

4.5.2 Simulation Study Evaluating the Approximate Bayesian Bootstrap

A simulation study is carried out to evaluate the performance of the ABB and the multiple imputation variance estimator. The case of predictive mean matching imputation is considered, however, not based on nearest neighbour imputation as in Heitjan and Little (1991) but for the case of hot deck imputation within classes where the classes are defined based on the predicted values of the imputation model. The ABB is as follows. The respondents in sample s are bootstrapped with replacement M times. The M samples are denoted $s_r^{(m)}$, $m=1,...,M$. The linear regression model underlying the imputation process is fitted in the M bootstrap samples and the vector of coefficients $\hat{\beta}^{(m)}$, $m=1,...,M$, is obtained for each sample $s_r^{(m)}$. The M sets of coefficients are used to obtain M sets of predicted values, which allow the generation of M independent multiple imputations. The imputation method now allows for the variation in the parameters. A similar approach is also used in Schenker and Taylor (1996), however, they draw the parameter estimates from the posterior distribution instead of using the ABB.

a.) Design of the Simulation Study

The design of the simulation study is as described in section 3.3. Nonresponse is introduced based on two different nonresponse mechanisms, uniform nonresponse and MAR nonresponse based on Model A3. Different response rates (rr) are considered for the uniform nonresponse mechanism, rr=0.43%, 60%, 70%, 80%. Having generated the sample $s^{(a)}$ for iteration a the respondents in sample $s^{(a)}$ are bootstrapped M times, sampling respondent units from $s^{(a)}$ with replacement. M samples are obtained, denoted $s_r^{(a,m)}$, where $m=1,...,M$. The imputation model is fitted in each of these samples and a set of estimated coefficients, denoted $\hat{\beta}^{(a,m)}$ for all $m=1,...,M$, is obtained. Based on each of the M sets of estimated coefficients M sets of predicted values, $\ln(\hat{y}_{regi}^{(a,m)})$, are obtained, where

$$\ln(\hat{y}_{regi}^{(a,m)}) = \eta_i^{(a)}\hat{\beta}^{(a,m)},\ \forall i \in s^{(a)},\ m=1,...,M. \tag{4.65}$$

Units $i \in s^{(a)}$ are allocated to imputation classes, denoted $B_k^{(m)}$ for $m=1,...,M$, based on each of the M sets of predicted values. The boundaries of the imputation classes are fixed as in section 3.3. and based on £1.5 pay bands. Within each class one donor is selected at random with replacement

from sample $s^{(a)}$ for each nonrespondent. Repeating this M times gives M imputed values for each nonrespondent in $s^{(a)}$. The results for the point estimator and the performance of the multiple imputation variance estimator are given in the following.

b.) Performance of Point Estimators

Table 4.7 shows the results for the two point estimators $\hat{P}_1.$ and $\hat{P}_2.$ under the ABB. Both estimators are approximately unbiased under uniform nonresponse. Under MAR nonresponse both estimators are significantly biased. However, the (estimated) relative bias is under 2%. This might be caused by the use of imputation classes as discussed in chapter 5. Using hot deck imputation with ABB gives, as expected, very similar results to hot deck imputation without ABB for both nonresponse mechanisms. For reasons of comparison the results for hot deck imputation without ABB are given in table 4.8. (The results are also reported in table 3.5.)

	Bias of $\hat{P}_1.$	Rel. Bias of $\hat{P}_1.$	Bias of $\hat{P}_2.$	Rel. Bias of $\hat{P}_2.$
uniform (rr= 43%)	$1.65*10^{-4}$ $(0.83*10^{-4})$	0.29 %	$-1.20*10^{-4}$ $(1.39*10^{-4})$	-0.06 %
MAR	$4.62*10^{-4}$ $(0.81*10^{-4})^*$	0.80 %	$29.82*10^{-4}$ $(1.50*10^{-4})^*$	1.61 %

Table 4.7: Performance of point estimators, $\hat{P}_1.$ and $\hat{P}_2.$, under hot deck imputation within classes with ABB, selecting donors with replacement (rr=response rate).

	Bias of $\hat{P}_1.$	Rel. Bias of $\hat{P}_1.$	Bias of $\hat{P}_2.$	Rel. Bias of $\hat{P}_2.$
uniform (rr= 43%)	$0.73*10^{-4}$ $(0.84*10^{-4})$	0.13 %	$2.59*10^{-4}$ $(1.42*10^{-4})$	0.13 %
MAR	$2.80*10^{-4}$ $(0.68*10^{-4})^*$	0.49 %	$26.21*10^{-4}$ $(1.53*10^{-4})^*$	1.41 %

Table 4.8: Performance of point estimators, $\hat{P}_1.$ and $\hat{P}_2.$, under hot deck imputation within classes without ABB, selecting donors with replacement.

c.) Performance of the Multiple Imputation Variance Formula

The aim is to analyse the performance of the multiple imputation variance estimator $\hat{var}_{MI}(\hat{P}_1.)$ using the ABB and to compare its properties to the performance of the variance formula, $\hat{var}_{(2)DRI}(\hat{P}_1.)$, previously derived for random hot deck imputation within classes without the ABB. As can be seen from the results in tables 4.2 and 4.3 for hot deck imputation without the ABB, the multiple imputation variance formula does not perform well with an (estimated) relative bias of around -30% and a coverage rate of approximately 90% under uniform nonresponse. The performance under MAR nonresponse is better with a relative bias of about -8% and a coverage rate of around 95%. As expected the formula $\hat{var}_{(2)DRI}(\hat{P}_1.)$ performs well under uniform nonresponse with a relative bias close to zero and a coverage rate of 95%. Under MAR nonresponse the relative bias is slightly higher with -2.8%. The coverage rate is close to 95%.

Introducing higher variability in the estimates by estimating the coefficients based on the bootstrap samples improves, as expected, the performance of $\hat{var}_{MI}(\hat{P}_1.)$. The results are presented in table 4.9 and 4.10. The relative bias under MAR nonresponse and under uniform nonresponse with a response rate of 80% is close to zero and the coverage rate is close to 95%. There is an indication that for MAR nonresponse the performance of $\hat{var}_{MI}(\hat{P}_1.)$ under the ABB might be slightly better than for $\hat{var}_{(2)DRI}(\hat{P}_1.)$ without the ABB when comparing the relative bias. However, the formula $\hat{var}_{MI}(\hat{P}_1.)$ does not perform well under uniform nonresponse with low response rates. In these cases the relative bias is between -9% and -22% and the coverage rate reduces to only 90% for a response rate of 43%. We also observe a larger simulation variance of the variance estimator for $\hat{var}_{MI}(\hat{P}_1.)$ under the ABB than for $\hat{var}_{(2)DRI}(\hat{P}_1.)$ without the ABB. The variance of the length of the 95% confidence interval is smaller for $\hat{var}_{(2)DRI}(\hat{P}_1.)$ without the ABB than for $\hat{var}_{MI}(\hat{P}_1.)$ with the ABB. The average length of the CI is approximately the same for both formulae (tables 4.3 and 4.10).

The fact that under uniform nonresponse with low nonresponse rates the performance of $\hat{var}_{MI}(\hat{P}_1.)$ is worse than under MAR nonresponse using the ABB might be related to the fact that under MAR nonresponse, although the overall response rate is 43% as in the uniform case, the response rate at the bottom end of the distribution of the variable Y, which is of interest for the estimators, is with between 70% and 80% much higher. Under uniform nonresponse with an

overall response rate of 43%, however, the response rate at the bottom end of the distribution is also only about 43%. The ABB does not seem to produce enough variability for low nonresponse rates to capture the increase in the nonresponse variability (V_{NR}). The formula $\hat{var}_{MI}(\hat{P}_1.)$ performs, however, well for moderate nonresponse rates. The performance of $\hat{var}_{MI}(\hat{P}_1.)$ under the ABB therefore seems to be related to the nonresponse rate. The derived variance formula $\hat{var}_{(2)DRJ}(\hat{P}_1.)$ performs well also under low response rates.

	Bias	Rel. Bias	RMSE	Rel. RMSE
uniform (rr= 43%)	$-14.91*10^{-7}$	-22.25 %	$17.10*10^{-7}$	25.55 %
uniform (rr= 60%)	$-7.22*10^{-7}$	-13.28 %	$9.20*10^{-7}$	16.90 %
uniform (rr= 70%)	$-4.31*10^{-7}$	-8.85 %	$6.25*10^{-7}$	12.82 %
uniform (rr= 80%)	$-0.04*10^{-7}$	-0.09 %	$3.03*10^{-7}$	7.31 %
MAR	$0.38*10^{-7}$	0.89 %	$3.85*10^{-7}$	8.92 %

Table 4.9: Simulation results for multiple imputation variance estimator, $\hat{var}_{MI}(\hat{P}_1.)$, under hot deck imputation with ABB, selecting donors with replacement.

	Simulation Variance of Variance Estimator	Rel. Simulation Variance of Variance Estimator	Average Length of CI	Variance of Length of CI	Coverage Rate
uniform (rr= 43%)	$6.66*10^{-13}$	$9.65*10^{-6}$ %	$9.08*10^{-3}$	$5.91*10^{-7}$	90.9 %
uniform (rr= 60%)	$3.24*10^{-13}$	$5.96*10^{-6}$ %	$8.57*10^{-3}$	$3.09*10^{-7}$	93.1 %
uniform (rr= 70%)	$2.04*10^{-13}$	$4.19*10^{-6}$ %	$8.30*10^{-3}$	$2.01*10^{-7}$	94.7 %
uniform (rr= 80%)	$0.91*10^{-13}$	$2.21*10^{-6}$ %	$7.99*10^{-3}$	$0.93*10^{-7}$	94.8 %
MAR	$1.53*10^{-13}$	$3.56*10^{-6}$ %	$8.20*10^{-3}$	$1.54*10^{-7}$	95.2 %

Table 4.10: Simulation results for multiple imputation variance estimator, $\hat{var}_{MI}(\hat{P}_1.)$ under hot deck imputation with replacement with ABB.

Under the ABB we also expect an increase in the simulation variance

$$V(\hat{\bar{P}}.) = \frac{1}{A-1}\sum_{a=1}^{A}(\hat{\bar{P}}.^{(a)} - \bar{P})^2 \tag{4.66}$$

where $\bar{P} = A^{-1}\sum_{a=1}^{A}\hat{\bar{P}}.^{(a)}$. The results for $V(\hat{\bar{P}}_1.)$ as well as the (estimated) mean square error $MSE(\hat{\bar{P}}_1.)$ are given in table 4.11. For hot deck imputation with the ABB the simulation variance $V(\hat{\bar{P}}_1.)$ shows an increase in comparison to without the ABB. For uniform nonresponse with a response rate of 43% the increase is 12% in comparison to without the ABB. Under MAR nonresponse the increase, however, is small, which implies that the additional variability introduced by using the ABB might be moderate.

Method	Nonresponse Mechanism	$V(\hat{\bar{P}}_1.)$	$MSE(\hat{\bar{P}}_1.)$
Hot Deck Imputation with ABB	uniform (rr=43%)	$7.65*10^{-6}$	$7.67*10^{-6}$
	uniform (rr=60%)	$5.44*10^{-6}$	$5.45*10^{-6}$
	uniform (rr=70%)	$4.87*10^{-6}$	$4.88*10^{-6}$
	uniform (rr=80%)	$4.14*10^{-6}$	$4.15*10^{-6}$
	MAR	$4.62*10^{-6}$	$4.67*10^{-6}$
Hot Deck Imputation without ABB	uniform (rr=43%)	$6.77*10^{-6}$	$6.82*10^{-6}$
	MAR	$4.58*10^{-6}$	$4.64*10^{-6}$

Table 4.11: Simulation variance $V(\hat{\bar{P}}_1.)$ and mean square error $MSE(\hat{\bar{P}}_1.)$ under uniform and MAR nonresponse.

We analyse the effects of the ABB on the estimation of the regression coefficients and the resulting predicted values that are used to allocate units to imputation classes in the simulation study. Although the ABB produces sets of coefficient estimates, $\hat{\beta}^{(a,m)}$, $m = 1,...,M$, based on the bootstrap samples that differ from $\hat{\beta}^{(a)}$, the estimated coefficients based on sample $s^{(a)}$, they do not seem to differ very much. The M sets of fitted values under the ABB are very close to the set of fitted values based on $\hat{\beta}^{(a)}$. We therefore have $\ln(\hat{y}_{regi}^{(a,m)}) \doteq \ln(\hat{y}_{regi}^{(a)})$ for $m = 1,...,M$ and $i \in s^{(a)}$. That means that the allocation of units to imputation classes for with and without ABB is

very similar. Further investigation shows that approximately only around 3.5% of all units in sample $s^{(a)}$ are allocated to different imputation classes using the ABB in comparison to without the ABB. One reason for $\hat{\beta}^{(a,m)} \doteq \hat{\beta}^{(a)}$ and therefore $\ln(\hat{y}_{reg\,i}^{(a,m)}) \doteq \ln(\hat{y}_{reg\,i}^{(a)})$ is believed to be due to the large sample size of n=15,000.

Another reason why the effect of the ABB is moderate might be related to the design of the simulation study. We investigated the effects of choosing narrower imputation classes resulting in an increase in the total number of classes (28 classes were defined instead of originally 7). It was hoped that the small differences in the predicted values for with and without replacement would have a greater impact. However, the effect was moderate such that the results are not reported here.

Another way of increasing the variability of the point estimates and to generate proper multiple imputation is to use the 'classical' approach of ABB for a non-parametric hot deck imputation within classes, which is not based on a regression model, as proposed by Rubin (1987). Having defined the imputation classes the respondents within each class are bootstrapped with replacement M times. For each nonrespondent a donor is selected at random from each set of the bootstrapped respondents. However, this approach was not followed up here.

We have seen that using the ABB improved the performance of the multiple imputation variance formula. However, the formula did not perform well for low nonresponse rates. The derived variance formula $\hat{\text{var}}_{(2)DRI}(\hat{P}_1.)$ performs well also under low nonresponse rates but showed a slightly higher bias than $\hat{\text{var}}_{MI}(\hat{P}_1.)$ under MAR nonresponse.

4.6 Conclusion

This chapter investigated the variance of the point estimator of interest taking into account sampling, nonresponse and imputation variability using a design-based (three phase) approach assuming uniform nonresponse within classes. It was shown that the variance can be decomposed into three components referring to sampling, nonresponse and imputation variability. An approximately unbiased variance estimator was derived, estimating the three components of the

variance separately. In addition, the performance of Rubin's multiple imputation formula was investigated. Because the imputation method does not constitute 'proper' multiple imputation, this approach was shown to underestimate the variance. It was demonstrated that the formula estimates the variance according to sampling and imputation correctly but underestimates the nonresponse variance. A modified multiple imputation formula was derived estimating the true variance correctly. Different variance estimators were evaluated in a simulation study showing good results for the newly derived variance estimators. As expected, the standard variance estimator valid under full response as well as Rubin's multiple imputation variance formula underestimate the true variance for the hot deck imputation method used here. A modified formula of the multiple imputation variance estimator showed good results in the simulation study. The newly derived variance formula $\hat{\text{var}}_{(2)DRI}(\hat{P}.)$ is recommended for practical use.

In addition, the approximate Bayesian bootstrap method was implemented and the variance estimator following Rubin's rule for variance estimation under proper multiple imputation was used. Under MAR nonresponse the formula performs well. However, under uniform nonresponse with low response rates (less than 60%) the formula shows a low coverage rate and a high relative bias. An indication of a larger variance of the estimator obtained using the ABB in comparison to using the improper multiple imputation method was found, which might make the point estimator under hot deck imputation within classes more efficient than under the proper multiple imputation method considered here.

Chapter 5

Comparing Predictive Mean Matching Imputation and Propensity Score Weighting

This chapter extends the previous imputation method and considers a wider class of adjustment methods for missing data. In section 5.1 various forms of predictive mean matching imputation are investigated and compared to the previous hot deck imputation within classes. In section 5.2 predictive mean matching imputation is compared to propensity score weighting and the close relationship between imputation and weighting is analysed. The main focus is on the empirical investigation of the properties of the estimators under the different approaches to missing data, using a simulation study. Some theoretical results are presented focussing on point estimation and the comparison between weighting and imputation. The aim is to find a method that produces approximately unbiased estimators with small variance and that is reasonably robust against misspecification of underlying models and assumptions. Some concluding remarks are given in section 5.3.

5.1 Predictive Mean Matching Imputation

5.1.1 Brief Theoretical Investigation of Predictive Mean Matching Imputation

In this section a class of imputation methods based on predictive mean matching is investigated. The basic form of predictive mean matching imputation is nearest neighbour imputation based on the predicted values. This method can be modified by applying repeated imputation or implementing a penalty function. Stochastic versus deterministic modifications are also explored. The previously described hot deck imputation method, also a form of predictive mean matching imputation, has the drawback that it relies on the specification of imputation classes, whose boundaries are more or less defined arbitrarily. A potential problem with the method of hot deck imputation within classes is that within imputation cells the (unknown) true value of Y for the nonrespondent and the imputed value from its donor might still be widely different even if the underlying model is correct since the classes might not be very narrowly defined. This might have an impact on the bias of the estimator. Nearest neighbour imputation addresses these problems by imputing the value of the "closest" respondent and defining essentially a class for each nonrespondent. The basic form of predictive mean matching imputation, as described in Little (1988, July), Heitjan and Little (1991) and Landerman, Land and Pieper (1997), is based on a regression model to obtain the predicted values for the variable Y for all units in the sample. According to these predicted values a distance D_{ji} is defined as in (1.45). For the LFS application considered here the regression model is given in (2.9) and the distance D_{ji} is defined as

$$D_{ji} = | \hat{y}_{reg\,j} - \hat{y}_{reg\,i} | \tag{5.1}$$

for each nonrespondent $j \in \bar{r}$ and all respondents $i \in r$. The value of Y from respondent i^* is chosen as the imputed value for nonrespondent j, such that the distance D is minimised, where

$$D_{ji^*} = \min_i | \hat{y}_{reg\,j} - \hat{y}_{reg\,i} | . \tag{5.2}$$

The imputed value is then $\hat{y}_j = y_{i^*}$. This approach has similarities to a sequential hot deck procedure, since the respondents and nonrespondents are essentially ordered according to their predicted value. Such a method is deterministic and represents a mixture of model-based and

nearest neighbour imputation or distance function matching. It has the advantage that it uses the donor who 'best' matches the nonrespondent in terms of the variables used to construct the distance measure. Therefore, finding the closest match may reduce the likelihood that within imputation classes the true value of Y and the imputed value differ significantly (Lessler and Kalsbeek, 1992, p. 218), which is a potential disadvantage of hot deck imputation within classes. The predictive mean matching method makes use of the regression predictor but is a semi-parametric method, since it does not fully rely on the specifications of the model. However, it achieves a considerable degree of conditioning on covariates given in the dataset, which according to Little (1988, July) is a desirable characteristic of any imputation method. It is a hot deck procedure and uses actual values for the imputation. Although nearest neighbour imputation is a commonly used imputation method only very recently have the theoretical properties been investigated. Rancourt, Särndal and Lee (1994) state that nearest neighbour imputation under the Euclidean distance function using just one auxiliary variable X gives approximately unbiased results for point estimates assuming a linear relationship between X and Y. Chen and Shao (2000) show that this is also true when nearly no assumption is made about the relationship between X and Y. Furthermore, they proved that under the assumption of MAR the nearest neighbour approach provides asymptotically unbiased and consistent estimators of functions of population means and totals, population distributions and population quantiles. Therefore, although nearest neighbour imputation is a deterministic method it is possible to estimate distributions correctly. To obtain approximately unbiased estimators under predictive mean matching imputation, based on the predicted values from the regression model (2.9), instead of the classical nearest neighbour imputation based on one covariate, we need to assume that the assumption of MAR holds with respect to the predicted values, i.e.

$$Y \perp I \mid Y_{reg}, \tag{5.3}$$

where Y_{reg} denotes the variable of the values y_{regi}, $y_{regi} = \exp(\eta_i \beta)$ is the exponential function of the predicted value $\eta_i \beta$ and η_i is a vector of functions of the derived variable X and other covariates. This assumption seems reasonable if MAR, as defined in assumption 2.2, holds. Since the nearest neighbour imputation method is defined with respect to $\hat{\beta}$ and assumption (5.3) is defined with respect to the vector β we need to assume that for large samples the estimator $\hat{\beta}$ converges to a vector β and that close neighbours with respect to the predicted values based on

$\hat{\beta}$, i.e. $\hat{Y}_{reg}$, are also close neighbours with respect to the predicted values based on β, Y_{reg}, which seems reasonable.

In addition to investigating the bias of estimators it is also important to consider the efficiency of point estimators under predictive mean matching imputation. Applying the method of predictive mean matching to LFS data may cause some of the donors to be used very often whereas others may not be used at all, since missing values occur much more frequently towards the top end of the distribution. There are several ways of avoiding such a multiple usage of donors and to achieve a smoothing of weights, which potentially leads to a reduction in variance of the estimator. One possibility is the use of a penalty function that can be included in the distance measure. The distance measure (5.2) is then modified to

$$D_{ji^*} = \min_i \{ | \hat{y}_{reg\,j} - \hat{y}_{reg\,i} | * (1 + \lambda t_i) \}, \tag{5.4}$$

where λ is a penalty factor, $\lambda \in \mathbb{R}^+$, and t_i is the number of times the respondent i has already been used as a donor (Kalton, 1983). Another way of spreading the usage of donors more evenly is the application of repeated imputations, also referred to as fractional imputation (Fay, 1996) or improper multiple imputation (Rubin, 1987).

In the following several variations of predictive mean matching are investigated. The method can be refined by choosing for example the nearest $M/2$ donors above and below the predicted value for the nonrespondent to obtain overall M imputed values. In our analysis we chose M=2 and M=10 repeated imputations. This predictive mean matching method can be modified to a stochastic method by defining $L/2$ nearest neighbours above and below the predicted value for the nonrespondent and by choosing M out of these L possible imputations at random. This can be done either with or without replacement. In our analysis we chose L=20, M=10 and L=4, M=2 and select the M donors out of the set of L possible donors with replacement. Another possibility to achieve a more equal use of donor values and therefore a reduction in variance is the previously discussed form of predictive mean matching imputation based on forming imputation classes defined by the range of the predicted values $\hat{Y}_{reg}$. Using hot deck imputation within classes is therefore expected to lead to a gain in efficiency in comparison to the basic form of nearest neighbour imputation based on the predicted values.

An approximate formula for the variance of a point estimator under nearest neighbour imputation is derived in Chen and Shao (2000), which is developed further for predictive mean matching imputation in Beissel-Durrant and Skinner (2003). The following conditional and unconditional expectations and variances are with respect to sampling design D and a superpopulation model ξ, assuming that $(y_i, \ln(y_{regi}), I_i)$ are iid random variables from a superpopulation and $y_i \perp I_i \mid \ln(y_{regi})$ for all $i \in U$. Let the function ψ be defined as $\psi(\ln(y_{regi})) = E(z_i \mid \ln(y_{regi}))$ and let d_{Mi} denote the imputation weights, which are defined as $d_{Mi} = 1 + (a_i / M)$, where a_i denotes the total number of times a respondent is used as a donor, and let r denote the set of respondents. Assuming n to be large and certain regularity conditions as given in Chen and Shao's theorem 1, an approximate variance formula of an imputed estimator under nearest neighbour imputation, using the argument of conditioning, is

$$\text{var}(\hat{P}.) \doteq \frac{1}{n^2} E[\sum_{i \in r} d_{Mi}^2 V(z_i \mid \ln(y_{regi}))] + \frac{1}{n} V[\psi(\ln(Y_{reg}))], \tag{5.5}$$

which is given in theorem 2 of Chen and Shao, equation (3.3). The variance formula in Chen and Shao (2000) is based on $M=1$ single value imputation but also holds for $M>1$ as used here.

We have

$$P(1-P) = \text{var}(Z) = E[V(Z \mid \ln(Y_{reg}))] + V[\psi(\ln(Y_{reg}))] \tag{5.6}$$

such that
$$V[\psi(\ln(Y_{reg}))] = P(1-P) - E[V(Z \mid \ln(Y_{reg}))]. \tag{5.7}$$

From theorem 1 in Chen and Shao (2000) it follows for large n that

$$E[\frac{1}{n} \sum_{i \in r} d_{Mi} V(z_i \mid \ln(y_{regi}))] \doteq E[V(Z \mid \ln(Y_{reg}))]. \tag{5.8}$$

Using (5.7) and (5.8) we can rewrite the variance formula in (5.5) in the following way

$$\text{var}(\hat{P}.) \doteq \frac{P(1-P)}{n} + \frac{1}{n^2} E[\sum_{i \in r} d_{Mi}^2 V(z_i \mid \ln(y_{regi}))] - \frac{1}{n^2} E[\sum_{i \in r} d_{Mi} V(z_i \mid \ln(y_{regi}))]$$

$$= \frac{P(1-P)}{n} + \frac{1}{n^2} E[\sum_{i \in r} d_{Mi}(d_{Mi} - 1) V(z_i \mid \ln(y_{regi}))], \tag{5.9}$$

which can be interpreted more easily. The first part of the variance formula (5.9) reflects the standard variance formula valid in case of complete response. The second term reflects additional variability due to imputation. We see that the inflation of the variance due to the second term depends on the imputation weights and therefore on the number of times a respondent is used as a donor for imputation. The same effect is discussed by David et al. (1986). To reduce the variance of the point estimator $\hat{P}.$ under predictive mean matching it is therefore of interest to reduce the variability in the imputation weights d_{Mi} and to achieve a smoothing of these weights. As discussed above this can be done by using modifications to the basic form of nearest neighbour imputation such as repeated imputation, a penalty function or imputation within classes. The use of donors and its effects on the variance of a point estimator under various forms of predictive mean matching are discussed in section 5.1.3. Variance estimation based on the formula (5.9) is briefly considered in Beissel-Durrant and Skinner (2003). Using the result from Theorem 1 of Chen and Shao (2000) that the imputed point estimator is approximately unbiased for the parameter P, it follows that

$$\hat{\text{var}}(\hat{P}.) = \frac{\hat{P}.(1-\hat{P}.)}{n} + \frac{1}{n^2}\sum_{i\in r} d_{Mi}(d_{Mi}-1)\hat{V}(z_i \mid \ln(y_{regi})) \tag{5.10}$$

is an approximately unbiased estimator of $\text{var}(\hat{P}.)$, if $\hat{V}(z_i \mid \ln(y_{regi}))$ is an approximately unbiased estimator of $V(z_i \mid \ln(y_{regi}))$. An estimator $\hat{V}(z_i \mid \ln(y_{regi}))$ therefore needs to be specified. Other variance estimation formulae for nearest neighbour imputation, for example using the Jackknife, have also been proposed (Rancourt, Saerndal and Lee 1994; Shao, 1997; Chen and Shao, 1999 and 2001; Rancourt 1999; Fay 1999).

5.1.2 Simulation Study Comparing Different Forms of Predictive Mean Matching Imputation

a.) Design of the Simulation Study

To investigate the properties of different forms of predictive mean matching imputation and the performances of the two point estimators of interest a simulation study is carried out. In each iteration a sample $s^{(a)}$, $a = 1,...,A$ of size n=15,000 employees is generated by selecting employees with replacement from one quarter of the LFS. Each sample $s^{(a)}$ is obtained in the

same way as described in section 3.3. In total A =1,000 iterations are used. Nonresponse is introduced using two different response mechanisms, uniform and MAR nonresponse. The latter is based on the logistic model A3 with nonlinear terms and interactions (see section 3.3.4 for details). The missing values are imputed using the following forms of predictive mean matching imputation.

a.) Nearest neighbour imputation based on the distance function (5.2) with one imputed value, M=1, referred to as NN1.

b.) Nearest neighbour imputation using the penalty function as defined in (5.4), with M=1. This method is denoted NN1p. Several values for λ were used. Only the results for $\lambda = 0.5$ are reported.

c.) Nearest neighbour imputation using M repeated imputations by taking $M/2$ donors above and below the predicted value for each nonrespondent, for M=2 and M=10, denoted NN2 and NN10 respectively.

d.) Nearest neighbour imputation using M repeated imputations by taking $L/2$ nearest neighbours above and below the predicted value for the nonrespondent and selecting M out of these L possible imputations at random, L=4, M=2 and L=20, M=10, selection with replacement, denoted NN2(4) and NN10(20) respectively.

In comparison to these different forms of nearest neighbour imputation the results for hot deck imputation within classes are reported, sampling donors with and without replacement using ten repeated imputations, denoted HDIwr10 and HDIwor10 respectively. In addition to these repeated or improper multiple imputation methods a proper multiple imputation method is also implemented using the Approximate Bayesian Bootstrap based on hot deck imputation within classes selecting ten multiple imputations with replacement, denoted ABB10(HDIwr) (see section 4.5). Bias and relative bias of the proportion of employees earning below the NMW, $\hat{P}_1.$, and the proportion of employees earning between the NMW and £5, $\hat{P}_2.$, are investigated. These two point estimators are chosen as estimators of the bottom end of the pay distribution and slightly higher up. In addition, the variance of the point estimators and the overall mean square error under the different methods are compared. When investigating the performance of the imputation

methods considered here the aim is to obtain a small mean square error, i.e. small bias and small variance.

b.) Bias and Relative Bias for Predictive Mean Matching Imputation

The (estimated) bias and relative bias of the two point estimators of interest under uniform and MAR nonresponse are given in tables 5.1 and 5.2. The relative biases for both proportions under all forms of nearest neighbour imputation are, with less than 1% for both nonresponse mechanisms, very small. Both single and repeated nearest neighbour imputation give very good results with approximately zero bias for both point estimators. Both the deterministic and random approaches of this imputation method perform well.

	Bias of $\hat{P}_1$.	Rel. Bias of $\hat{P}_1$.	Bias of $\hat{P}_2$.	Rel. Bias of $\hat{P}_2$.
NN1	$0.3*10^{-4}$ $(1.1*10^{-4})$	0.06 %	$3.1*10^{-4}$ $(2.0*10^{-4})$	0.16 %
NN1p[1]	$7.1*10^{-4}$ $(3.3*10^{-4})^*$	1.24%	$-1.3*10^{-4}$ $(5.2*10^{-4})$	-0.07 %
NN2	$1.9*10^{-4}$ $(1.1*10^{-4})$	0.33 %	$2.7*10^{-4}$ $(1.8*10^{-4})$	-0.15 %
NN2(4)	$2.8*10^{-4}$ $(1.2*10^{-4})^*$	0.51 %	$-0.5*10^{-4}$ $(1.8*10^{-4})$	-0.02 %
NN10	$1.8*10^{-4}$ $(1.0*10^{-4})$	0.32 %	$0.4*10^{-4}$ $(1.6*10^{-4})$	0.02 %
NN10(20)	$-0.1*10^{-4}$ $(1.0*10^{-4})$	-0.02 %	$0.6*10^{-4}$ $(1.7*10^{-4})$	0.03 %
HDIwr10	$0.7*10^{-4}$ $(0.8*10^{-4})$	0.13 %	$2.5*10^{-4}$ $(1.4*10^{-4})$	0.13 %
HDIwor10	$1.7*10^{-4}$ $(0.8*10^{-4})$	0.31 %	$-0.9*10^{-4}$ $(1.5*10^{-4})$	-0.05 %
ABB10(HDIwr)	$1.7*10^{-4}$ $(0.8*10^{-4})$	0.29 %	$-1.2*10^{-4}$ $(1.4*10^{-4})$	-0.06 %

Table 5.1: Bias and relative bias for both point estimators for various forms of predictive mean matching imputation under uniform nonresponse. (The numbers in brackets show the standard error of the bias, $\sqrt{V}/\sqrt{A}$, $V=(A-1)^{-1}\sum_{a=1}^{A}(\hat{P}_{\cdot}^{(a)}-\bar{P})^2$ and A the number of iterations. A star (*) indicates that the bias is significantly different from zero on a 95% significance level.)

	Bias of $\hat{P}_1$.	Rel. Bias of $\hat{P}_1$.	Bias of $\hat{P}_2$.	Rel. Bias of $\hat{P}_2$.
NN1	$1.2*10^{-4}$ $(0.9*10^{-4})$	0.22 %	$0.9*10^{-4}$ $(1.7*10^{-4})$	0.04 %
NN1p[2]	$4.4*10^{-4}$ $(2.6*10^{-4})$	0.78 %	$0.3*10^{-4}$ $(5.1*10^{-4})$	0.01 %
NN2	$0.6*10^{-4}$ $(8.5*10^{-4})$	0.10 %	$1.6*10^{-4}$ $(1.5*10^{-4})$	0.08 %
NN2(4)	$1.4*10^{-4}$ $(0.9*10^{-4})$	0.25 %	$-2.5*10^{-4}$ $(1.5*10^{-4})$	-0.13 %
NN10	$0.2*10^{-4}$ $(6.5*10^{-4})$	0.04 %	$-1.2*10^{-4}$ $(1.5*10^{-4})$	-0.06 %
NN10(20)	$0.2*10^{-4}$ $(0.8*10^{-4})$	0.03 %	$0.7*10^{-4}$ $(1.5*10^{-4})$	0.04 %
HDIwr10	$2.8*10^{-4}$ $(0.7*10^{-4})^*$	0.49 %	$26.2*10^{-4}$ $(1.5*10^{-4})^*$	1.41 %
HDIwor10	$2.5*10^{-4}$ $(0.7*10^{-4})^*$	0.45 %	$28.1*10^{-4}$ $(1.2*10^{-4})^*$	1.55 %
ABB10(HDIwr)	$4.6*10^{-4}$ $(0.8*10^{-4})^*$	0.80 %	$29.8*10^{-4}$ $(1.5*10^{-4})^*$	1.61 %

Table 5.2: Bias and relative bias for both point estimators for various forms of predictive mean matching imputation under MAR nonresponse. (A star (*) indicates that the bias is significantly different from zero on a 95% significance level.)

Only in the case of imputation choosing 2 out of 4 donor values or nearest neighbour with a penalty function both under uniform nonresponse is the bias significantly different from zero. In all other cases the bias is not significantly different from zero for the different forms of nearest neighbour imputation considered here. For hot deck imputation within classes (HDIwr10, HDIwor10 and ABB10(HDIwr)) the biases of the two point estimators under uniform nonresponse are also not significant.

Under MAR nonresponse, however, the biases for both point estimators under these three methods are significantly different from zero, although the biases are very small. This suggests that

[1] Note: due to computing time only A =100 iterations were used.

the methods based on imputation classes can potentially lead to biased results, which might be caused by the choice and width of the imputation classes. It is expected that smaller classes may lead to less biased results. We conclude that the different forms of nearest neighbour imputation considered here perform well under the conditions of the simulation study. There is evidence that nearest neighbour imputation might perform better than the hot deck imputation within classes with respect to the bias. (Note that the results in table 5.1 and 5.2 for HDI are presented in chapter 3 table 3.5 and 3.6 and the results for ABB10 are presented in chapter 4 table 4.7).

c.) Comparison of Efficiency for Different Forms of Predictive Mean Matching Imputation

It is of interest to compare the efficiency of the point estimators under the different predictive mean matching methods considered here. In the following the simulation variance V of the two point estimators under each method is investigated, which is defined as

$$V(\hat{\bar{P}}_g) = \frac{1}{A-1}\sum_{a=1}^{A}(\hat{\bar{P}}_g^{(a)} - \bar{P}_g)^2, \tag{5.11}$$

where $\bar{P}_g = A^{-1}\sum_{a=1}^{A}\hat{\bar{P}}_g^{(a)}$ and $g = 1,2$. The aim is to obtain a small variance V. Table 5.3 and table 5.4 show the results for the variance, the efficiency gain when using a certain imputation method in comparison to nearest neighbour imputation using only one imputation (NN1), i.e. the ratio V/V_{NN1}, and the mean square error for both point estimators under uniform and MAR nonresponse. Since the conclusions from the results under uniform or MAR nonresponse are similar we concentrate on the results for MAR nonresponse. As expected the variance V is largest for nearest neighbour with one imputation (NN1). Under nearest neighbour imputation using repeated imputations the variance is reduced in comparison to single imputation with a greater reduction as M increases. The variance is smallest for hot deck imputation within classes in particular for the case of selecting donors without replacement. Analysing the ratio V/V_{NN1} we can see that the efficiency gain for using the penalty function NN1p as supposed to NN1 is about 10%. For M=2 repeated imputations we also observe a gain in efficiency of about 10% in

[2] Note: due to computing time only A =100 iterations were used.

comparison to the basic form of nearest neighbour imputation. The greatest reduction in variance is achieved by using M=10 repeated imputations with an efficiency gain of about 20% in comparison to NN1. Based on the simulation results the greatest reduction in variance is achieved by hot deck imputation without replacement with more than 20%. No obvious advantages between deterministic and random predictive mean matching methods, such as between NN2 and NN2(4) as well as NN10 and NN10(20), were found. With respect to the mean square error hot deck imputation within classes with and without replacement (HDIwr10, HDIwor10) and the nearest neighbour with 10 and 10 out of 20 imputations (NN10, NN10(20)) show the best performances. Overall the differences between the mean square errors under the various methods are small. We conclude that repeated imputation has appreciable gains in efficiency in comparison to single imputation. Taking into account the results for both bias and variance of the point estimators nearest neighbour imputation using repeated imputations, such as NN10, performs best under the conditions of the simulation study.

Method	$V(\hat{P}_1.)$	$V(\hat{P}_2.)$	$\frac{V(\hat{P}_1.)}{V_{NN1}(\hat{P}_1.)}$	$\frac{V(\hat{P}_2.)}{V_{NN1}(\hat{P}_2.)}$	$MSE(\hat{P}_1.)$	$MSE(\hat{P}_2.)$
NN1	$1.39*10^{-5}$	$4.07*10^{-5}$	1	1	$1.39*10^{-5}$	$4.07*10^{-5}$
NN1p[3]	$1.12*10^{-5}$	$3.55*10^{-5}$	0.80	0.87	$1.16*10^{-5}$	$3.55*10^{-5}$
NN2	$1.25*10^{-5}$	$3.31*10^{-5}$	0.89	0.81	$1.20*10^{-5}$	$3.31*10^{-5}$
NN2(4)	$1.15*10^{-5}$	$3.38*10^{-5}$	0.82	0.83	$1.15*10^{-5}$	$3.38*10^{-5}$
NN10	$1.06*10^{-5}$	$2.85*10^{-5}$	0.76	0.70	$1.06*10^{-5}$	$2.85*10^{-5}$
NN10(20)	$1.06*10^{-5}$	$3.23*10^{-5}$	0.76	0.79	$1.06*10^{-5}$	$3.23*10^{-5}$
HDIwr10	$1.10*10^{-5}$	$3.12*10^{-5}$	0.79	0.76	$1.10*10^{-5}$	$3.12*10^{-5}$
HDIwor10	$1.05*10^{-5}$	$2.78*10^{-5}$	0.75	0.69	$1.05*10^{-5}$	$2.79*10^{-5}$
ABB10(HDIwr)	$1.14*10^{-5}$	$3.28*10^{-5}$	0.82	0.79	$1.14*10^{-5}$	$3.28*10^{-5}$

Table 5.3: Variance V, ratio of variance V to reference variance, which is the variance for NN1, and mean square error for both point estimators under uniform nonresponse for different predictive mean matching imputation methods.

[3] Note: due to computing time only A =100 iterations were used.

Method	$V(\hat{P}_1.)$	$V(\hat{P}_2.)$	$\frac{V(\hat{P}_1.)}{V_{NN1}(\hat{P}_1.)}$	$\frac{V(\hat{P}_2.)}{V_{NN1}(\hat{P}_2.)}$	$MSE(\hat{P}_1.)$	$MSE(\hat{P}_2.)$
NN1	$7.81*10^{-6}$	$2.95*10^{-5}$	1	1	$7.86*10^{-6}$	$2.95*10^{-5}$
NN1p[4]	$6.82*10^{-6}$	$2.66*10^{-5}$	0.87	0.91	$7.05*10^{-6}$	$2.66*10^{-5}$
NN2	$7.23*10^{-6}$	$2.55*10^{-5}$	0.91	0.86	$7.20*10^{-6}$	$2.55*10^{-5}$
NN2(4)	$7.52*10^{-6}$	$2.38*10^{-5}$	0.94	0.80	$7.51*10^{-6}$	$2.38*10^{-5}$
NN10	$6.53*10^{-6}$	$2.38*10^{-5}$	0.83	0.81	$6.57*10^{-6}$	$2.38*10^{-5}$
NN10(20)	$6.61*10^{-6}$	$2.30*10^{-5}$	0.84	0.77	$6.61*10^{-6}$	$2.30*10^{-5}$
HDIwr10	$6.40*10^{-6}$	$2.17*10^{-5}$	0.82	0.74	$6.47*10^{-6}$	$2.86*10^{-5}$
HDIwor10	$6.20*10^{-6}$	$2.23*10^{-5}$	0.78	0.76	$6.26*10^{-6}$	$3.02*10^{-5}$
ABB10(HDIwr)	$6.91*10^{-6}$	$2.37*10^{-5}$	0.88	0.80	$7.10*10^{-6}$	$3.25*10^{-5}$

Table 5.4: Variance V, ratio of variance V to reference variance, which is the variance for NN1, and mean square error for both point estimators under MAR nonresponse for different predictive mean matching imputation methods.

d.) Robustness against Model Misspecification

Since predictive mean matching imputation incorporates the predictive regression model based on the derived variable and other covariates it is of interest to analyse the effects of model misspecification on the performance of this imputation method. In general, it is expected that the method of predictive mean matching as a form of semi-parametric regression imputation is less sensitive to model misspecification than choosing a method that explicitly uses the regression model for the imputation process such as parametric regression imputation (Little, 1988, July; Schenker and Taylor, 1996). Landerman, Land and Pieper (1991) investigate the impact of different regression models and conclude that the predictive mean matching method performs reasonably well under misspecification, if the model has sufficient predictive power. Also Sande (1988) emphasizes the importance of the predictive efficiency when using such a method. We therefore investigate several possible regression models for the imputation process that differ from the models used to generate the data. In the following it is assumed that the model specified in section 2.2.4, including all significant variables, nonlinear terms and interactions describes the

[4] Note: due to computing time only A =100 iterations were used.

relationship between $\ln(Y)$ and other variables best. We therefore generate $\ln(Y)$ in the simulation study according to this model. Linear regression models used in the imputation procedure that deviate from this model reflect misspecification. The imputation method investigated here is nearest neighbour with ten repeated imputations (NN10). Note that the results for HDI under model misspecification are reported in section 3.4.

The results under model misspecification are given in table 5.5. We find that using a regression model with only $\ln(X)$ or $\ln(X)$ and $\ln(X)^2$ results in a high relative bias of the point estimator of around 10%. A regression model that includes only the main effects of the significant variables but excludes the variable $\ln(X)$ shows a relative bias of around 7%. Using the model with the variable $\ln(X)$ but only some of the significant variables, (e.g. $\ln(X)$, $\ln(X)^2$, major occupation group (SOCcat), qualification (Qcat), AGE, AGE^2, industry section (INCcat), regions (REGcat) and gender) results in a small relative bias of 1.39%. The regression model that incorporates all of the significant main effects including the variable $\ln(X)$ but no nonlinear terms or interactions, performs very well with a relative bias of 0.31%. As expected, the best performance is given by the regression model with all of the significant main effects, the nonlinear terms and the interactions with no significant bias.

We conclude that very simple linear regression models, such as including only the variable $\ln(X)$ do not show very satisfactory results. The imputation method seems to be sensitive to violent model misspecification. However, very good results are obtained for all models that include the variable $\ln(X)$ and at least some additional explanatory variables, with or without nonlinear terms or interactions. We therefore conclude that the proposed method seems to be robust against minor model misspecification. In comparison to hot deck imputation within classes the sensitivity to model misspecification shows a very similar pattern. However, the performance seems slightly worse for the hot deck method with for example a relative bias of approximately 9% and 3% for the two models 'including all main effects apart from $\ln(X)$' and 'only some of the significant variables including $\ln(X)$' respectively (see table 3.9).

Explanatory Variables used in Linear Regression Model	Bias of $\hat{P}_1$.	Rel. Bias of $\hat{P}_1$.
$\ln(X)$	$56.8*10^{-4}$ $(2.0*10^{-4})^*$	10.02 %
$\ln(X)+\ln(X)^2$	$56.0*10^{-4}$ $(1.9*10^{-4})^*$	9.95 %
all main effects apart from $\ln(X)$	$39.2*10^{-4}$ $(1.5*10^{-4})^*$	6.91 %
some of the sign. variables including $\ln(X)$	$7.9*10^{-4}$ $(0.9*10^{-4})^*$	1.39 %
all main effects including $\ln(X)$ but no nonlinear terms and no interactions	$1.7*10^{-4}$ $(0.8*10^{-4})^*$	0.31%
main effects, nonlinear terms and interactions	$0.2*10^{-4}$ $(0.8*10^{-4})$	0.04 %

Table 5.5: Impact of model misspecification under MAR nonresponse for nearest neighbour with ten repeated imputations (NN10). (A star (*) indicates that the bias is significantly different from zero on a 95% significance level.)

5.1.3 Analysis of the Use of Donors for Different Nearest Neighbour Imputation Methods

As discussed in section 5.1.1 it is desirable to spread the use of donors or to limit the number of times a donor is used since this can lead to a reduction in variance due to a smoothing in the imputation weights. We therefore investigate how often a donor is used for imputation under the different nearest neighbour imputation methods. Of interest is also how many of the respondents are not used in the imputation process. This is relevant to both variance and robustness. We may expect the variance of the estimates to be less the less variable the number of times that different donors are used. In addition, we may have robustness concerns about donors which are used a very large number of times, since the resulting estimates may be over-dependent upon the value of the direct variable for such donors. Since the counts of how often a donor is used for imputation are difficult to compare for the various methods, due to the different choices of M, we also

consider the differences in the variance of imputation weights. The following results have been obtained by applying NN1, NN1p, NN2 and NN10 to the LFS quarter March-May 2000.

Under hot deck imputation within classes without replacement the use of donors is per definition approximately evenly spread within each imputation class. We therefore concentrate on the usage of donors among the various forms of nearest neighbour imputation. The results are given in table 5.6. Under nearest neighbour with one imputed value (NN1) the usage of donors is very concentrated with about 51% of respondents not used. A very small proportion of donors is used very often, i.e. between 25 and 60 times. Incorporating the penalty function relaxes this concentration, however, not dramatically with about 41% of donors not used at all. This might be because only few missing values occur at the bottom end of the hourly pay distribution whereas many missing values occur towards the upper end. It was found that the number of respondents not used decreases, as expected, with increasing value of λ. The donors not used for imputation, mainly occur in the bottom end of the distribution, whereas higher up all donors are used at least once. Imputing two values under nearest neighbour imputation approximately 35% of respondents are not used. For M=10 imputations (NN10) only 4% of donors were not used at all. However, 11% of respondents are used very often, i.e. between 25 and 250 times.

Percentage of Donors	NN1	NN1p	NN2	NN10
not used	51.1	41.7	35.5	4.1
used once	23.5	32.4	22.4	6.9
used twice	10.6	12.2	12.6	9.2
used >25 times	0.1	0.06	1.2	11.5

Table 5.6: Percentage of donors that are not used, used once, twice or more than 25 times under different imputation methods.

Since in the LFS there is a higher proportion of respondents at the bottom end of the hourly pay distribution, the use of donors strongly depends on the level of hourly pay. The percentage of donors used is therefore investigated within deciles of the level of hourly pay. The deciles are defined based on the predicted values from the regression model for imputation. The cut-off points for the deciles are given in table 5.7.

Decile	Log(cut-off)	Cut-off
1	1.4493	£4.22
2	1.5565	£4.71
3	1.6656	£5.25
4	1.7753	£5.87
5	1.8832	£6.55
6	2.0133	£7.46
7	2.1656	£8.67
8	2.3669	£10.59
9	2.6691	£14.42

Table 5.7: Cut-off points for the deciles based on the predicted values from the regression model.

	NN1	NN1p	NN2	NN10
Decile 1				
not used in dec 1	78.3	76.1	63.6	14.8
used once in dec 1	18.5	22.0	26.1	20.8
used twice in dec 1	2.5	1.7	6.5	20.0
used 3 times in dec 1	0.5	0.2	2.5	16.4
used >3 times in dec 1	0.2	0	1.1	27.8
Decile 2				
not used in dec 2	75.0	71.7	57.7	7.2
used once in dec 2	19.4	25.2	27.8	14.6
used twice in dec 2	4.5	2.6	9.9	23.1
used 3 times in dec 2	0.8	0.4	2.8	18.7
used >3 times in dec 2	0.2	0	1.7	36.6
Decile 3				
not used in dec 3	58.5	50.1	39.4	0.6
used once in dec 3	28.0	40.9	27.8	3.2
used twice in dec 3	8.9	7.7	16.2	7.5
used 3 times in dec 3	3.5	1.0	9.5	10.8
used >3 times in dec 3	3.5	0.1	7.1	77.7

Table 5.8: Percentage of donors (based on the number of donors in each decile) that are not used, used once, twice, three times or more than three times within the first, second and third deciles.

Table 5.8 shows the percentage use of donors within deciles. We can see that in the bottom deciles a large proportion of donor values are not used for imputation which is expected since the response rate is higher in this decile. This proportion is highest for nearest neighbour (NN1) imputation and lowest for NN10 imputation. The proportion of donors used more than three times is by far the highest for NN10 for the bottom three deciles.

However, the analysis of the percentage of donors used is only an indication of the effective usage of donors. The different methods are difficult to compare since nearest neighbour based on repeated imputations per definition uses donors more frequently than single value imputation. We therefore analyse the variance of the imputation weights d_{Mi} to be able to compare different nearest neighbour methods directly. The variance of the imputation weights is defined as

$$\text{var}(d_{Mi}) = \frac{1}{n_r - 1}\sum_{i\in r}(d_{Mi} - \bar{d}_M)^2 \tag{5.12}$$

where $\bar{d}_M$ denotes the mean of the imputation weights, $\bar{d}_M = n_r^{-1}\sum_{i\in r} d_{Mi} = n/n_r$. The variance of the point estimator under predictive mean matching imputation given in (5.9) depends on the weights d_{Mi} and is therefore inflated by a factor, which depends upon the variance of the imputation weights. The variance $\text{var}(\hat{P}.)$ is large if the variance $\text{var}(d_{Mi})$ is large and small if $\text{var}(d_{Mi})$ is small. The results for $\text{var}(d_{Mi})$ for NN1, NN1p, NN2 and NN10 are given in table 5.9. We can see that $\text{var}(d_{Mi})$ reduces if M increases (M=1,2,10). The results in table 5.9 suggest that the use of repeated imputation, in particular M=10, may be expected to reduce the variance of the resulting estimator significantly. The use of the penalty function also appears to have a substantial effect on variance reduction. In addition, table 5.10 gives the results for the variance of the imputation weights within the bottom three deciles of the hourly pay distribution based on the predicted values for $\ln(Y)$, where $\text{var}(d_{Mi})$ for the c-th decile dec_c, $c = 1,...,10$, is defined as

$$\text{var}_c(d_{Mi}) = \frac{1}{[\sum_i I(i \in \{r \cap dec_c\})] - 1}\sum_{i\in r\cap dec_c}(d_{Mi} - \bar{d}_M)^2. \tag{5.13}$$

Again we observe that the variance of the imputation weights is greatly reduced using NN10 instead of single imputation. We conclude that NN10 spreads the use of donors more evenly which potentially leads to a reduction in the variance of estimators.

NN1	NN1p	NN2	NN10
7.88	5.58	7.35	4.59

Table 5.9: Variance of the imputation weights, $\text{var}(d_{Mi})$, for different forms of nearest neighbour imputation.

Decile	NN1	NN1p	NN2	NN10
1	1.40	1.34	1.28	1.15
2	1.36	1.28	1.20	1.03
3	1.23	0.97	0.95	0.59

Table 5.10: Variance of the imputation weights within the bottom three deciles, $\text{var}_c(d_{Mi})$, c=1,2,3.

5.2 Propensity Score Weighting and Comparison to Predictive Mean Matching Imputation

5.2.1 Brief Theoretical Investigation of Propensity Score Weighting

An alternative way of handling nonresponse is weighting, which is usually applied in the case of unit nonresponse. The method implies weighting the respondents to compensate for nonresponse bias and either dropping nonrespondents from the file or assigning zero weights to the nonrespondents. In this case the estimator $\hat{P}_.$ can be written as a weighted estimator, such that

$$\hat{P}_. = \frac{\sum_{i \in r} \omega_i z_i}{\sum_{i \in r} \omega_i}, \tag{5.14}$$

where ω_i denotes the weights for the respondents to compensate for nonresponse bias. For simplicity reasons the survey weights w_i are not taken into account in the following. There are various ways of obtaining the adjusted weights ω_i for the respondents. One method of forming weights is based on the response propensity. The respondents' weights are set proportional to the inverse of response rates. This way it is possible to adjust for selective nonresponse. In the following the method based on the inverse of the response rates will be referred to as *propensity score weighting* (David et al., 1983; Agostino and Rubin, 2000; Little and Rubin, 2002), denoted PSW. The approach can be regarded as analogous to using the inverse of the sampling probabilities for compensating for differential sample selection, such as in the Horvitz-Thompson estimator (Oh and Scheuren, 1983). Instead of estimating the hourly pay distribution $f(y_i \mid x_i, \mathrm{w}_i, I_i = 0)$ directly, which reduces under the assumption of MAR to estimating the distribution $f(y_i \mid x_i, \mathrm{w}_i, I_i = 1)$, as it is done under imputation, the aim when using propensity score weighting is to estimate the distribution $Pr(I_i = 1 \mid y_i, x_i, \mathrm{w}_i)$, which under MAR reduces to estimating $p_i = Pr(I_i = 1 \mid x_i, \mathrm{w}_i)$. This distribution is then used to estimate the distribution of hourly pay. The model $Pr(I_i = 1 \mid x_i, \mathrm{w}_i)$ can be fitted based on observed data only since I, X and W are fully observed. A regression model is fitted with the indicator of response I as the outcome variable and X and W defining the covariates using for example a logistic regression model to obtain the predicted *response propensities* $\hat{p}_i = \hat{Pr}(I_i = 1 \mid x_i, \mathrm{w}_i)$ for all $i \in s$. The weights for each respondent in s are set to be $\omega_i = \hat{p}_i^{-1}$. The resulting estimator is denoted $\hat{P}_{.PSW}$ and is of the form

$$\hat{P}_{.PSW} = \frac{\sum_{i \in r} \hat{p}_i^{-1} z_i}{\sum_{i \in r} \hat{p}_i^{-1}}. \qquad (5.15)$$

Note that $E_R(\sum_{i \in r} \hat{p}_i^{-1}) \doteq n$, where the expectation is with respect to the nonresponse mechanism R. The simplest approach to propensity score weighting is to use only the *response rate*, defined as n_r / n, for weighting the respondents. Alternatively, classes, denoted B_k, $k = 1, ..., K$, can be formed based on boundaries of the predicted propensity scores $\hat{p}_i$. Then within each class the response rate, n_{rk} / n_k, is used for weighting the respondents within class k. This is referred to as *response propensity stratification* (David et al., 1983). Another way of defining the classes is to use categorical variables that are observed for all $i \in s$.

Under certain conditions, which will be examined in the following, propensity score weighting leads to approximately unbiased estimators. First, it is required that the assumption of MAR holds with respect to $p_i = Pr(I_i = 1 | x_i, \mathrm{w}_i)$. Rosenbaum and Rubin (1983) (see also David et al., 1983) show that for the propensity score p_i it holds

$$(x_i, \mathrm{w}_i) \perp I_i \mid p_i \tag{5.16}$$

Under MAR it follows that

$$y_i \perp I_i \mid p_i . \tag{5.17}$$

Using equation (5.17) the estimator $\hat{P}_{.PSW}$ is approximately unbiased assuming that the underlying propensity score model is correctly specified (David et al. 1983). More formally, we have

$$E_{DR}(\hat{P}_{.PSW}) = E_{DR}\left(\frac{\sum_{i \in s} \hat{p}_i^{-1} I(i \in r) \mathrm{z}_i}{\sum_{i \in s} \hat{p}_i^{-1} I(i \in r)}\right) \doteq P ,$$

since $E_D(\hat{p}_i) = p_i$ if $\hat{\varphi}$ converges to a limit φ, where $\hat{p}_i = \hat{Pr}(I_i = 1 | x_i, \mathrm{w}_i) = h[\eta_i \hat{\varphi}]$, $p_i = Pr(I_i = 1 | x_i, \mathrm{w}_i) = h[\eta_i \varphi]$, η is a vector of functions of X and covariates W and φ is a vector of coefficients estimated by $\hat{\varphi}$. If π_i denotes the true probability of response for $i \in s$, i.e. $\pi_i = Pr(I_i = 1 | y_i, x_i, \mathrm{w}_i)$, we have per definition $E_R(I(i \in r)) = \pi_i$. Under MAR it follows that $p_i = \pi_i$, which implies approximate unbiasedness of the estimator.

5.2.2 Comparison of Propensity Score Weighting to Predictive Mean Matching Imputation

The method of propensity score weighting in the context of estimating the hourly pay distribution and the parameter P, has been proposed by Dickens and Manning (2002) and Manning and Dickens (2002). It is therefore of interest to compare the properties of this weighting method with the previously discussed forms of predictive mean matching imputation and to investigate advantages and disadvantages of the two approaches, particularly in the context of deriving estimates of low pay. As examined in Little (1988, July) and Oh and Scheuren (1983) a close relationship between weighting and certain forms of imputation, such as hot deck imputation,

exists. In the case of predictive mean matching imputation, i.e. hot deck imputation within classes and nearest neighbour imputation, it is possible to express the point estimator $\hat{P}.$ as a weighted estimator as in (5.14). The weights ω_i are defined by the imputation weights, i.e. $\omega_i = d_{Mi}$, where $d_{Mi} = 1 + a_i / M$ and a_i denotes the number of times respondent $i \in r$ has been used as a donor for imputation (see section 5.1.1). Note that per definition $\sum_{i \in r} d_{Mi} = n$. The weighted estimator (5.14) under hot deck imputation is denoted $\hat{P}_{.IMP}$, such that

$$\hat{P}_{.IMP,M} = \frac{\sum_{i \in r} d_{Mi} z_i}{\sum_{i \in r} d_{Mi}}. \tag{5.18}$$

If the estimator (5.18) is derived using hot deck imputation within classes it will be denoted $\hat{P}_{.HDI,M}$ and if it is based on nearest neighbour imputation it will be denoted $\hat{P}_{.NN,M}$. Expressing the imputed estimator as a weighted estimator might have potential advantages for users of LFS data. The use of improper multiple imputation, in particular in the case of HDI10 and NN10, might cause problems to users of the data when combining the results of the M imputed variables. However, the close relationship between imputation and weighting might help to overcome this problem since the variable representing the imputation weights could be included in the dataset instead of the M repeated imputation variables. Estimates of population values can be obtained by weighting the respondents data appropriately using the estimator in (5.18) instead of combining the results based on each separate imputation. The variable of the imputation weights is also sufficient when computing estimates of the variance for the point estimators of interest, if variance estimation formulae are expressed based on the imputation weights, which is done in Beissel-Durrant and Skinner (2003). The actual M imputed variables are then not necessarily needed.

Since both imputation and weighting can be used for nonresponse adjustment it is important to understand the theoretical properties and the relative merits of each approach. Comparisons between weighting and imputation have been analysed in greater detail, for example, by Little (1986), David et al. (1983) and Kuk, Mak and Li (2001). In the following the relationship between propensity score weighting and predictive mean matching imputation is briefly examined. We concentrate on a comparison between propensity score weighting and hot deck imputation within classes, since for a meaningful comparison we need the requirement that M is large or infinite,

which is not well defined for nearest neighbour imputation. Taking into account that the propensity score method is a deterministic method whereas the hot deck method within classes is a stochastic method given the sample s, we want to eliminate the stochastic element in the estimator (5.18) for hot deck imputation within classes. We therefore assume that M tends to infinity such that

$$\hat{P}_{\cdot HDI} = \frac{\sum_{i \in r} d_i^* z_i}{\sum_{i \in r} d_i^*}, \tag{5.19}$$

where $d_i^* = \lim_{M \to \infty} (d_{Mi}) = \lim_{M \to \infty} (1 + a_i / M)$.

First, some conditions are discussed under which the two methods produce similar estimates. Under MAR using hot deck imputation within classes and propensity score stratification, using the response rate within classes to define the weights, both methods are expected to give similar results if the classes for both methods are defined equivalently, for example by using the categories of some categorical covariates or by using an appropriate range of $\hat{p}_i$ or $\hat{y}_{regi}$ to define the classes respectively for each method. In the same way this also applies to using propensity score weighting instead of stratification under uniform within classes nonresponse. Assuming that the classes B_k, $k = 1,...,K$, are defined equivalently for both methods we have

$$\hat{P}_{\cdot PSW} = \frac{\sum_k \sum_{i \in B_k \cap r} \frac{n_k}{n_{rk}} z_i}{\sum_k \sum_{i \in B_k \cap r} \frac{n_k}{n_{rk}}} = \frac{1}{n} \sum_k \frac{n_k}{n_{rk}} \sum_{i \in B_k \cap r} z_i = \frac{1}{n} \sum_k n_k \bar{z}_{rk}, \tag{5.20}$$

since under response propensity stratification the response rate is used for weighting, i.e. $\hat{p}_i = n_{rk} / n_k$ for all $i \in B_k$, and

$$\hat{P}_{\cdot HDI} = \frac{1}{n} \sum_{i \in r} d_i^* z_i = \frac{1}{n} \sum_k \sum_{i \in B_k \cap r} (1 + a_{ik}^*) z_i = \frac{1}{n} \sum_k \frac{n_k}{n_{rk}} \sum_{i \in B_k \cap r} z_i = \frac{1}{n} \sum_k n_k \bar{z}_{rk}, \tag{5.21}$$

where $a_{ik}^* = \lim_{M \to \infty} (a_i / M) = \frac{n_{mk}}{n_{rk}} = \frac{n_k - n_{rk}}{n_{rk}} = \frac{n_k}{n_{rk}} - 1$.

This shows that under the above conditions the estimators for imputation and weighting are approximately the same. In general, however, the classes defined under hot deck and under response propensity weighting based on the range of $\hat{y}_{regi}$ or $\hat{p}_i$ respectively are not assumed to be the same since, for example, some covariates might be good predictors of Y but independent of R. Using the estimator $\hat{P}_{.HDI}$ but defining the classes based on response propensity stratification or vice versa will lead to biased results (Little, 1986).

The idea of defining equivalent classes and using propensity score stratification and hot deck imputation within classes extends to the more general case of using propensity score weighting and predictive mean matching imputation. Both methods are expected to produce close estimates if the following holds. The predicted values from the linear regression used in the imputation method are defined by $y_{regi} = \exp(\eta_i\beta)$, ignoring the difference in $\hat{\beta}$ and β. The propensity score can be written as $p_i = h[\eta_i\varphi]$, where the function h may be defined as $h(x) = \exp(x)/[1+\exp(x)]$ for example, and we denote $y_{regi}^{PSW} = \eta_i\varphi$, ignoring the difference in $\hat{\varphi}$ and φ. Both methods therefore depend on the specifications of the vectors of coefficients β and φ. If these are functions of each other, such that $\beta = f(\varphi)$, where f is a one-to one function imputation and propensity score weighting are expected to produce very similar results. One example is to set $f(x) = gx$, where $g \in \mathbb{R}$, such that $\beta = g\varphi$. It follows that

$$\ln(y_{regi}) = \eta_i\beta = \eta_i g\varphi = g\eta_i\varphi = g y_{regi}^{PSW}$$

and

$$p_i = h[\eta_i\varphi] = h[g^{-1}\eta_i\beta] = h[g^{-1}\ln(y_{regi})],$$

such that the predicted values from the imputation regression are functions of the propensity scores and vice versa. The simplest case would be to set $\beta = \varphi$, which means that the propensity scores are defined conditional on the predicted values from the linear regression model for imputation. We have $p_i = h[\eta_i\beta] = h[\ln(y_{regi})]$ and therefore $p_i = \Pr(I_i = 1 | \ln(y_{regi}))$. Then, the imputation weight d_{Mi} may be regarded as a nonparametric estimate of $\Pr(I_i = 1 | \ln(y_{regi}))^{-1}$. In this case, propensity score weighting and imputation are expected to lead to similar results. Generally, however, it does not hold that $\beta = f(\varphi)$ since some variables are expected to be predictors of hourly pay but not necessarily of the nonresponse and vice versa. The following simple example illustrates that given the sample s the actual estimators $\hat{P}_{.HDI}$ and $\hat{P}_{.PSW}$ are

different in general. Let us assume uniform nonresponse and only one categorical variable W that is a predictor of Y, but independent of I. The propensity score weights are therefore constant for all $i \in s$ and the estimator $\hat{P}_{\cdot PSW}$ reduces to

$$\hat{P}_{\cdot PSW} \doteq \frac{\sum_{i \in r} \frac{n}{n_r} z_i}{\sum_{i \in r} \frac{n}{n_r}} = \frac{1}{n_r} \sum_{i \in r} z_i \,. \tag{5.22}$$

Using hot deck imputation within classes, where the classes are either based on the range of the predicted values of Y or on the categories of W, we have

$$\hat{P}_{\cdot HDI} = \frac{1}{n} \sum_{k} \sum_{i \in B_k \cap r} \frac{n_k}{n_{rk}} z_i = \frac{1}{n} \sum_{k} n_k \bar{z}_{rk} \,, \tag{5.23}$$

which is unequal to (5.22).

We conclude that under MAR and correct model specifications both estimators for imputation and propensity score weighting are approximately unbiased for P. Furthermore, under certain conditions as discussed above both estimators may lead to very similar estimates. In general, however, the estimates are expected to differ since both methods are based on quite different models.

Another issue when comparing imputation and weighting is their reliance on model specifications and their robustness to misspecification of the underlying assumptions. Since propensity score weighting is a parametric method, which is based on a model for the response, the choice of variables employed in the regression model is important and the performance of the method depends on the quality of the underlying model (Little, 1988, July). In comparison, predictive mean matching imputation makes use of the underlying regression model but does not fully rely on it. It is a semi-parametric method and this might imply greater robustness in the case of model misspecification. This is discussed further in section 5.2.4, where both methods are analysed empirically under model misspecification.

Another important aspect when comparing imputation and weighting is the analysis of the efficiency of the resulting estimators. Although the emphasis is on the empirical investigation

using a simulation study as done in section 5.2.4 we will briefly discuss the efficiency under propensity score weighting and imputation. Similar to imputation standard variance formulae under weighting are not applicable and correct standard errors are difficult to obtain. An adjustment is necessary and is discussed by Jones and Chromy (1982) and Oh and Scheuren (1983). A variance formula for propensity score weighting is also derived in Beissel-Durrant and Skinner (2003) linearising the point estimator $\hat{P}_{\cdot PSW}$. The variance formula is given by

$$\text{var}(\hat{P}_{\cdot PSW}) \doteq \frac{P(1-P)}{n} + \frac{1}{n^2} E[\sum_{i \in r} \frac{1}{p_i}(\frac{1}{p_i} - 1) V(z_i \mid y_{regi}^{PSW})]$$

$$+ \frac{1}{n} E\{\sum_{i \in r}[\frac{1}{p_i} - 1][E(z_i \mid y_{regi}^{PSW}) - P]\}, \qquad (5.24)$$

where the unconditional and conditional variances and expectations are with respect to the sampling design and the superpopulation. We can see that this formula involves a third term in comparison to the variance formula under predictive mean matching imputation given in (5.9). As in this formula the inflation of the variance due to nonresponse in the formula in (5.24), as indicated by the second and third term, depends on the weights, here the propensity score weights $1/p_i$. We will therefore focus on the weights under imputation and weighting. As discussed in Little (1986) and David et al. (1983) propensity score weighting can lead to estimates with large variances since for very low values of p_i the assigned weights are large, which leads to an inflation of the variance of $\hat{P}_{\cdot PSW}$. However, this effect might be moderate in this particular application of estimating low pay since under an approximately correct model we would expect only high earners to receive a very small probability of response. For high earners, however, we would generally have $z_i = 0$ for proportions related to the NMW, such that the influence of large weights on the variance formula would be reduced. In comparison, one might expect to achieve a smoothing of the weights under predictive mean matching imputation if M is reasonably large, producing weights with smaller variability. An empirical analysis of the imputation weights and the weights under propensity score weighting is carried out in section 5.2.3 showing a higher dispersion of the propensity score weights in comparison to the imputation weights. In general, however, it is difficult to compare the variance of an estimator under propensity score weighting in (5.24) with

the variance under predictive mean matching imputation in (5.9) since the weights $1/\hat{p}_i$ and d_{Mi}, the underlying regression models and therefore y_{regi}^{PSW} and $\ln(y_{regi})$ are different.

5.2.3 Comparison of Propensity Score Weights and Imputation Weights

Comparing the estimators based on imputation and weighting it is of interest to analyse the distribution of the resulting weights empirically. The weights ω_i in the estimator (5.14) using propensity score weighting are defined as $\hat{p}_i^{-1} = \hat{\Pr}^{-1}(I_i = 1 | x_i, w_i)$. For hot deck imputation within classes as well as nearest neighbour imputation we have $\omega_i = d_{Mi} = 1 + a_i / M$. Table 5.11 gives the distribution of the resulting weights for the different methods applied to the LFS quarter March-May 2000 to get an indication of the differences of the weights. The variance of the weights is defined as

$$\text{var}(\omega_i) = \frac{1}{n_r - 1} \sum_{i \in r} (\omega_i - \bar{\omega})^2, \tag{5.25}$$

where $\bar{\omega} = n_r^{-1} \sum_{i \in r} \omega_i$. The variance $\text{var}(\omega_i)$ for the c-th decile dec_c, $c = 1,...,10$, where the deciles are based on the direct variable Y, is defined as

$$\text{var}_c(\omega_i) = \frac{1}{[\sum_i I(i \in \{r \cap dec_c\})] - 1} \sum_{i \in r \cap dec_c} (\omega_i - \bar{\omega})^2. \tag{5.26}$$

Table 5.11 presents the distribution of the weights ω_i based on propensity score weighting, nearest neighbour imputation and hot deck imputation within classes. Note that the underlying models, i.e. imputation model and propensity score model, both employ the variables as in the nonresponse model A3 (section 3.3.4) for comparison. The analysis might therefore favour the propensity score weighting method. We can see that the maximum value differs greatly with the greatest value for PSW and the second largest value for NN10. For hot deck imputation within classes without replacement (HDIwor10) the use of donors is as expected very evenly spread resulting in an even distribution of weights, which can be seen in the cut-off points for the deciles of the distribution of the weights. In table 5.11 we can see that the cut-off points for the deciles for propensity score weighting and nearest neighbour imputation are very similar. However, we observe a great difference in the variances of the weights, with the largest variance for the propensity score weighting method and the lowest variance for hot deck imputation within classes

without replacement. The variance of the weights within deciles of the observed variable Y show for nearly all deciles a higher variance for propensity score weighting than for nearest neighbour imputation based on ten repeated imputations with the largest discrepancy for the top deciles. This coincides with larger variances for the point estimators under propensity score weighting observed in the simulation study (see table 5.13).

	PSW	NN10	HDIwor10
Minimum	1.02	1	2
Maximum	109.91	28.12	3
Median	1.55	1.61	2.30
Mean	2.31	2.31	2.31
Deciles (Cut-off points):			
1	1.1	1.1	2.1
2	1.2	1.2	2.2
3	1.3	1.3	2.2
4	1.4	1.5	2.3
5	1.6	1.6	2.3
6	1.7	1.9	2.3
7	2.1	2.3	2.4
8	2.6	2.8	2.4
9	3.8	3.8	2.5
Variance	10.01	4.37	0.03
Variance of weights within deciles of the observed variable Y			
Decile 1	1.28	0.85	0.032
Decile 2	1.20	1.14	0.029
Decile 3	1.05	1.16	0.027
Decile 4	0.91	0.96	0.026
Decile 5	0.84	0.79	0.026
Decile 6	0.81	0.47	0.029
Decile 7	0.72	0.28	0.026
Decile 8	1.19	0.57	0.028
Decile 9	4.44	2.75	0.029
Decile 10	118.88	47.07	0.028

Table 5.11: Distribution of the weights ω_i under propensity score weighting (PSW), nearest neighbour imputation (NN10) and hot deck imputation within classes (HDIwor10) based on an application to the LFS quarter March-May 2000. The underlying model is model A3 for all three methods.

The variance of the weights within deciles is very small for hot deck imputation within classes without replacement. These findings support the arguments in section 5.2.2. The higher variability

in the propensity score weights is expected to lead to a higher variance of the point estimator under propensity score weighting in comparison to the imputation methods analysed here.

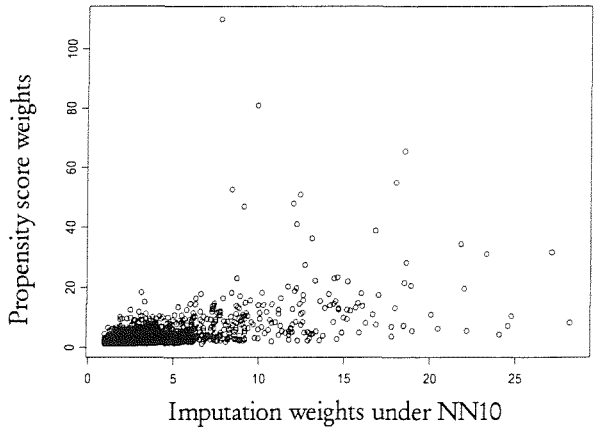

Figure 5.1: Scatterplot of imputation weights under NN10 and propensity score weights based on LFS quarter MM00, including all cases. The underlying model is model A3 for both methods.

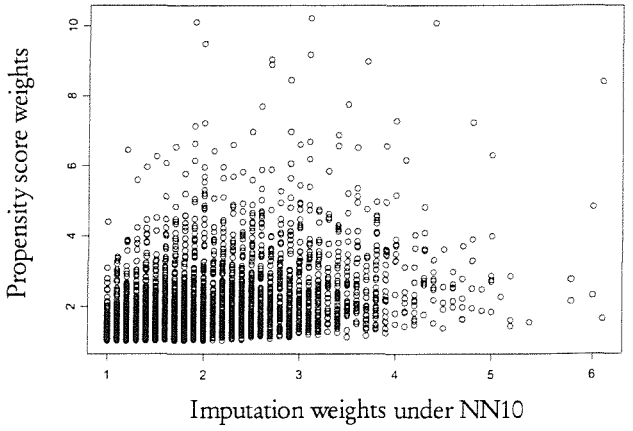

Figure 5.2: Scatterplot of imputation weights under NN10 and propensity score weights based on LFS quarter MM00, including the first 6000 cases (out of 6813), where the cases are ordered according to the predicted values of the imputation model used for NN10. The underlying model is model A3 for both methods.

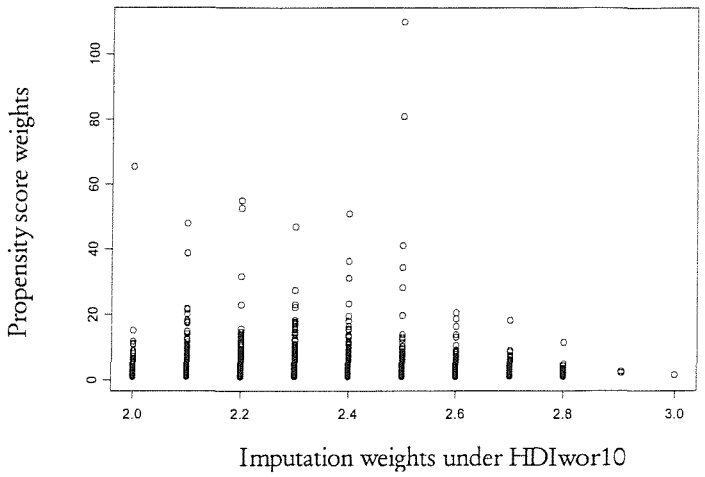

Figure 5.3: Scatterplot of imputation weights from HDIwor10 and propensity score weights based on LFS quarter MM00. The underlying model is model A3 for both methods.

The joint distribution of the imputation weights for NN10 and the propensity score weights is shown in figure 5.1 including all cases and in figure 5.2 including the first 6000 cases only (out of a total of 6813), where the cases are ordered according to the predicted values of the imputation model. It can be seen that the weights under NN10 can be quite different from the weights under propensity score weighting for each $i \in r$. We do not observe a linear relationship. Figure 5.3 presents the joint distribution of the imputation weights for HDIwor10 and the propensity score weights. As expected the weights for HDIwor10 take only on values within a short range, i.e. values between 2 and 3, reflecting a high degree of smoothing, whereas we observe a much wider range of possible values with some very large weights under propensity score weighting (values between 1 and 100).

5.2.4 Simulation Study Comparing Propensity Score Weighting and Predictive Mean Matching Imputation

The properties of propensity score weighting, nearest neighbour imputation and hot deck imputation within classes are compared empirically using a simulation study. Only NN10 and HDIwor10 are considered here. We concentrate on the properties of two point estimators, $\hat{P}_1.$, the proportion of employees earning below the NMW, and $\hat{P}_2.$, the proportion of employees

earning between the NMW and £5, in terms of bias, variance and mean square error. Furthermore, the sensitivity to model misspecification under imputation and weighting is examined.

a.) Design of the Simulation Study

The following simulation study is designed as in section 3.3, however, with the following amendments. Bias comparisons between predictive mean matching imputation methods based on the predictive distribution and the propensity score method using a propensity score model are difficult since the bias depends on the conditions of the simulation study and the specifications of the models involved. A comparison requires that the models for the two approaches are comparable. In the following investigation the same variables are employed for the imputation models and the propensity score model to ensure comparability. The nonresponse follows the assumption of MAR and is generated using Model A3. The variables included in this model are described in section 3.3.4. The linear regression model that generates the variable Y contains the same variables as in model A3. The variable $\ln(X)$ is generated as in section 3.3. The behaviour of the methods under ideal conditions and under misspecification of the models involved is analysed, i.e. the imputation model and the propensity score model depart from the model generating nonresponse and the model generating the variable Y in the simulation. The following variables are considered for modelling the propensity score as well as the imputation model. In addition to models that were found relevant in section 3.3.4 four models that are considered in the paper by Dickens and Manning (2002) are also taken into account, denoted 'models DM'.

1. Model A3: includes all covariates as given in table 3.4 including interactions and square terms
2. Model A1: includes all covariates as in Model A3 but omits interactions and square terms
3. Model A0: is a simpler model using less covariates including: ln(derived variable), occupation (SOC), part time (PT), less than weekly (LTWK), length of time continuously employed (EMPMON), additions to basic pay (ADDTBP), proxy response (Proxy), gender (FEMALE), industry section (IND), qualification (Q).
4. Model DM1: includes the covariates age, education (Qcat), gender (FEMALE), race (ETHNICDM), temporary worker (JOBTYPE), part time (PT), working in public sector (PUBLIC) and employer size (SIZE).
5. Model DM2: all variables as in model DM1 plus industry (IND) and occupation (SOC)
6. Model DM3: all variables as in model DM2 plus proxy response (PROXY)

7. Model DM4: all variables as in model DM3 plus ln(derived variable)

b.) Results of the Simulation Study

First, the (estimated) bias for the estimators under different model specifications is analysed. The results are given in table 5.12. The estimator $\hat{P}_1$. is approximately unbiased under weighting and imputation under the ideal condition using covariates as in model A3 for the underlying imputation and weighting models. However, if the propensity score model and the imputation model diverge from the ideal model conditions bias is introduced and the bias is greater the more these models diverge from the model generating the data and the model generating the nonresponse. The estimator $\hat{P}_1$. shows a greater bias under the propensity score method than under the two imputation methods for models A1, A0 and DM4, with nearest neighbour imputation (NN10) having the smallest bias. For models DM1, DM2 and DM3 all methods are strongly biased, since the derived variable is omitted and only a small number of covariates are taken into account. In these cases the propensity score method shows a slightly smaller bias than hot deck imputation within classes. The nearest neighbour approach gives the lowest bias under these misspecifications. Similar results are shown for the second point estimator $\hat{P}_2$.. For models A1, A0 and DM4 the second point estimator under the PSW method shows a higher relative bias with around 4% than under the imputation methods with under 2%, where the nearest neighbour method performs best showing the smallest bias. For models DM1, DM2 and DM3 the results for the second point estimator are all strongly biased, showing the smallest bias for the nearest neighbour imputation method. We conclude that in this comparison nearest neighbour imputation (NN10) seems to perform best in terms of bias under correct and misspecified models. Kuk, Mak and Li (2001) also found that under model misspecification imputation seems to perform better than weighting. This might be related to the stronger dependence of the weighting method as a parametric method on the specification of the underlying model.

Method	Model	Bias of $\hat{P}_1$.	Rel. Bias of $\hat{P}_1$.	Bias of $\hat{P}_2$.	Rel. Bias of $\hat{P}_2$.
PSW	A3	$0.14*10^{-4}$ $(0.71*10^{-4})$	0.03 %	$-2.62*10^{-4}$ $(1.35*10^{-4})$	-0.14 %
	A1	$-8.96*10^{-4}$ $(0.68*10^{-4})^*$	-1.64 %	$70.21*10^{-4}$ $(1.40*10^{-4})^*$	3.80 %
	A0	$-5.02*10^{-4}$ $(0.68*10^{-4})^*$	-0.92 %	$67.81*10^{-4}$ $(1.41*10^{-4})^*$	3.66 %
	DM1	$118.13*10^{-4}$ $(0.87*10^{-4})^*$	21.81 %	$428.54*10^{-4}$ $(1.58*10^{-4})^*$	23.17 %
	DM2	$48.7*10^{-4}$ $(0.79*10^{-4})^*$	8.95 %	$191.01*10^{-4}$ $(1.40*10^{-4})^*$	10.33 %
	DM3	$56.4*10^{-4}$ $(0.80*10^{-4})^*$	10.35 %	$205.12*10^{-4}$ $(1.47*10^{-4})^*$	11.13 %
	DM4	$-3.28*10^{-4}$ $(0.68*10^{-4})^*$	-0.60 %	$89.71*10^{-4}$ $(1.36*10^{-4})^*$	4.31 %
NN10	A3	$-0.18*10^{-4}$ $(0.64*10^{-4})$	-0.03 %	$-5.8*10^{-4}$ $(1.20*10^{-4})^*$	-0.31 %
	A1	$-1.31*10^{-4}$ $(0.65*10^{-4})^*$	-0.24 %	$-4.74*10^{-4}$ $(1.23*10^{-4})^*$	-0.25 %
	A0	$-1.66*10^{-4}$ $(0.63*10^{-4})^*$	-0.30 %	$-10.61*10^{-4}$ $(1.23*10^{-4})^*$	-0.57 %
	DM1	$117.11*10^{-4}$ $(0.85*10^{-4})^*$	21.54 %	$393.45*10^{-4}$ $(1.58*10^{-4})^*$	21.30 %
	DM2	$43.21*10^{-4}$ $(0.77*10^{-4})^*$	7.93 %	$154.34*10^{-4}$ $(1.32*10^{-4})^*$	8.37 %
	DM3	$46.01*10^{-4}$ $(0.76*10^{-4})^*$	8.45 %	$163.23*10^{-4}$ $(1.39*10^{-4})^*$	8.83 %
	DM4	$3.05*10^{-4}$ $(0.65*10^{-4})^*$	0.56 %	$26.30*10^{-4}$ $(1.24*10^{-4})^*$	0.14 %
HDIwor10	A3	$-1.1*10^{-4}$ $(0.64*10^{-4})$	-0.20%	$22.12*10^{-4}$ $(1.23*10^{-4})^*$	1.21 %
	A1	$-2.37*10^{-4}$ $(0.66*10^{-4})^*$	-0.43 %	$24.01*10^{-4}$ $(1.24*10^{-4})^*$	1.30 %
	A0	$-2.73*10^{-4}$ $(0.65*10^{-4})^*$	-0.50 %	$18.79*10^{-4}$ $(1.24*10^{-4})^*$	1.01 %
	DM1	$137.43*10^{-4}$ $(0.87*10^{-4})^*$	25.09 %	$448.45*10^{-4}$ $(1.57*10^{-4})^*$	24.25 %
	DM2	$55.8*10^{-4}$ $(0.78*10^{-4})^*$	10.25 %	$193.78*10^{-4}$ $(1.36*10^{-4})^*$	10.45 %
	DM3	$59.31*10^{-4}$ $(0.83*10^{-4})^*$	10.89 %	$205.90*10^{-4}$ $(1.42*10^{-4})^*$	10.95 %
	DM4	$1.64*10^{-4}$ $(0.65*10^{-4})^*$	0.30 %	$31.21*10^{-4}$ $(1.21*10^{-4})^*$	1.69 %

Table 5.12: Comparison of bias and relative bias under propensity score weighting (PSW), hot deck imputation within classes (HDIwor10) and nearest neighbour imputation (NN10) under comparable conditions, using the same covariates in the propensity score model as in the imputation models. The nonresponse mechanism follows the assumption of MAR and is based on model A3. The model specifies which variables were used in the PSW method and the imputation methods.

Method	Model	$V(\hat{P}_1.)$	$MSE(\hat{P}_1.)$	$V(\hat{P}_2.)$	$MSE(\hat{P}_2.)$
PSW	A3	$5.16*10^{-6}$	$5.16*10^{-6}$	$1.83*10^{-5}$	$1.83*10^{-5}$
	A1	$4.71*10^{-6}$	$5.51*10^{-6}$	$1.96*10^{-5}$	$6.90*10^{-5}$
	A0	$4.68*10^{-6}$	$4.94*10^{-6}$	$1.97*10^{-5}$	$6.59*10^{-5}$
	DM1	$7.66*10^{-6}$	$148.01*10^{-6}$	$2.52*10^{-5}$	$185.23*10^{-5}$
	DM2	$6.33*10^{-6}$	$30.19*10^{-6}$	$1.96*10^{-5}$	$38.44*10^{-5}$
	DM3	$6.44*10^{-6}$	$3.82*10^{-6}$	$2.18*10^{-5}$	$44.5*10^{-5}$
	DM4	$4.72*10^{-6}$	$4.83*10^{-6}$	$1.85*10^{-5}$	$5.09*10^{-5}$
NN10	A3	$4.10*10^{-6}$	$4.10*10^{-6}$	$1.45*10^{-5}$	$1.49*10^{-5}$
	A1	$4.27*10^{-6}$	$4.29*10^{-6}$	$1.51*10^{-5}$	$1.54*10^{-5}$
	A0	$4.08*10^{-6}$	$4.10*10^{-6}$	$1.52*10^{-5}$	$1.63*10^{-5}$
	DM1	$7.66*10^{-6}$	$145.34*10^{-6}$	$2.52*10^{-5}$	$157.23*10^{-5}$
	DM2	$6.05*10^{-6}$	$24.71*10^{-6}$	$1.76*10^{-5}$	$25.7*10^{-5}$
	DM3	$5.80*10^{-6}$	$26.97*10^{-6}$	$1.94*10^{-5}$	$28.55*10^{-5}$
	DM4	$4.27*10^{-6}$	$4.36*10^{-6}$	$1.55*10^{-5}$	$1.56*10^{-5}$
HDIwor10	A3	$4.20*10^{-6}$	$4.21*10^{-6}$	$1.51*10^{-5}$	$2.01*10^{-5}$
	A1	$4.38*10^{-6}$	$4.44*10^{-6}$	$1.54*10^{-5}$	$2.12*10^{-5}$
	A0	$4.24*10^{-6}$	$4.31*10^{-6}$	$1.56*10^{-5}$	$1.91*10^{-5}$
	DM1	$7.52*10^{-6}$	$195.55*10^{-6}$	$2.51*10^{-5}$	$203.34*10^{-5}$
	DM2	$6.19*10^{-6}$	$374.23*10^{-6}$	$1.87*10^{-5}$	$39.13*10^{-5}$
	DM3	$5.98*10^{-6}$	$41.14*10^{-6}$	$2.03*10^{-5}$	$43.02*10^{-5}$
	DM4	$4.43*10^{-6}$	$4.46*10^{-6}$	$1.57*10^{-5}$	$2.56*10^{-5}$

Table 5.13: Comparison of variance and mean square error under propensity score weighting (PSW), hot deck imputation within classes (HDIwr10) and nearest neighbour imputation (NN10) under comparable conditions, using the same covariates in the propensity score model as in the imputation models. The nonresponse mechanism follows the assumption of MAR and is based on model A3. The model specifies which variables were used in the PSW method and the imputation methods.

Table 5.13 shows the variance and the mean square error for the two estimators of interest. The efficiencies of the estimators are investigated based on the simulation variance $V=(A-1)^{-1}\sum_{a=1}^{A}(\hat{P}_{g\cdot}^{(a)}-\bar{P}_g)^2$, where $\bar{P}_g=A^{-1}\sum_{a=1}^{A}\hat{P}_{g\cdot}^{(a)}$ and $g=1,2$. Comparing the efficiency of $\hat{P}_{1\cdot}$ for the different methods we can see that under hot deck imputation within classes the values for $V(\hat{P}_{1\cdot})$ for all models considered here show lower values than under the propensity score method. For all models apart from DM1 and DM2 the nearest neighbour approach shows the smallest variance amongst the three methods. This indicates that, under the conditions considered here, the imputation methods based on repeated imputations, in particular nearest neighbour with 10 imputed values, are more efficient than propensity score weighting. The variance for the second point estimator is also smallest for nearest neighbour imputation for all models considered here. In addition, the mean square error is the smallest for nearest neighbour for all models and both point estimators. We can therefore conclude that overall nearest neighbour imputation based on repeated imputation seems to perform better in terms of bias and efficiency than both other methods considered here.

In addition, in section 6.2.2 both propensity score weighting and predictive mean matching imputation (HDI10 and NN10) are analysed under misspecification of the nonresponse mechanism, where the nonresponse is nonignorable rather than following the MAR assumption. The results indicate a slightly smaller relative bias for nearest neighbour imputation in comparison to hot deck imputation within classes and propensity score weighting. We can see that also under misspecification of the nonresponse mechanism there is an indication that nearest neighbour imputation seems to perform better than the other methods.

c.) Analysis of Point Estimators under Propensity Score Weighting and Predictive Mean Matching Imputation

The relationship between the three estimators $\hat{P}_{\cdot PSW}$, $\hat{P}_{\cdot HDI10}$ and $\hat{P}_{\cdot NN10}$ is investigated empirically. In particular, it is of interest if the estimators under weighting and imputation lead to similar results under comparable conditions, i.e. if $\hat{P}_{\cdot PSW}^{(a)}\doteq\hat{P}_{\cdot HDI10}^{(a)}\doteq\hat{P}_{\cdot NN10}^{(a)}$ for all iterations $a=1,...,A$, where the model generating $\ln(Y)$ in the simulation, the nonresponse model, the imputation model and the propensity score model employ the same variables.

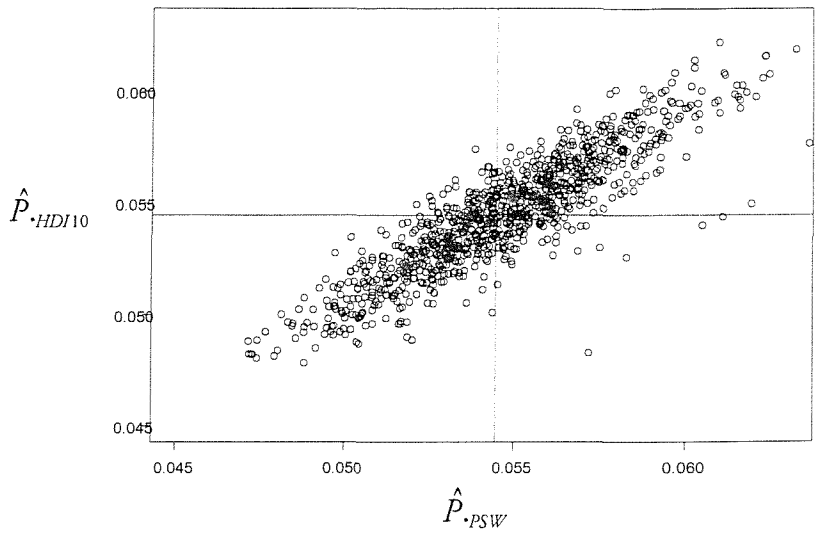

Figure 5.4: Joint distribution of $\hat{P}_{\cdot HDI10}$ and $\hat{P}_{\cdot PSW}$ under model A3. The two lines indicate the population value P.

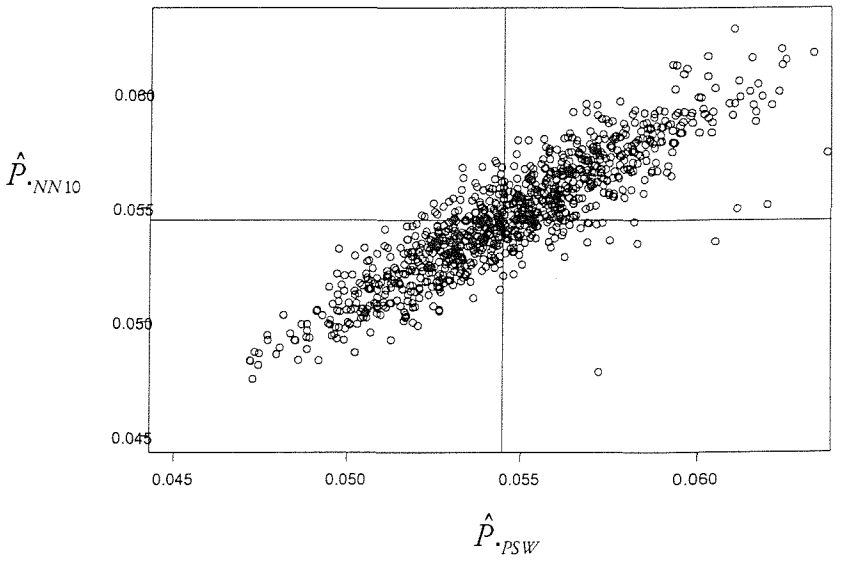

Figure 5.5: Joint distribution of $\hat{P}_{\cdot NN10}$ and $\hat{P}_{\cdot PSW}$ under model A3. The two lines indicate the population value P.

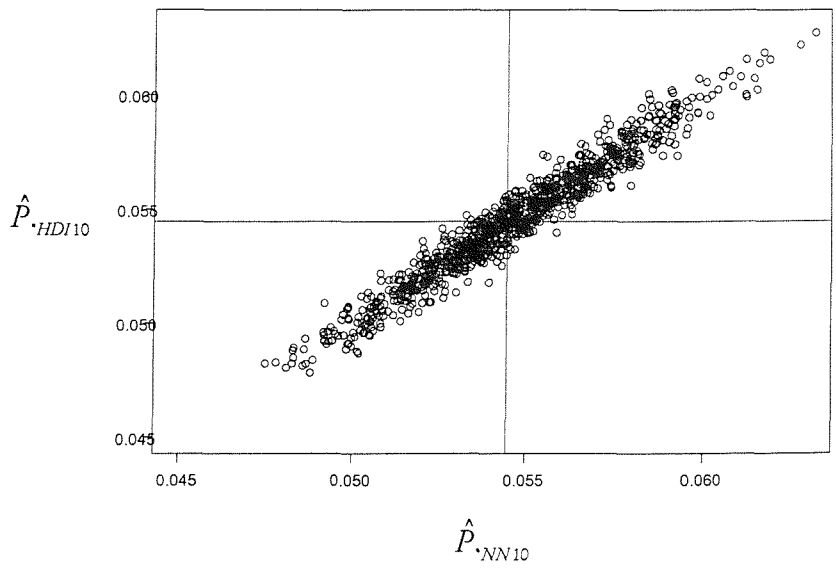

Figure 5.6: Joint distribution of $\hat{P}_{\cdot HDI10}$ and $\hat{P}_{\cdot NN10}$ under model A3. The two lines indicate the population value P.

Figures 5.4-5.6 show the joint distributions of $\hat{P}_{1\cdot HDI10}$, $\hat{P}_{1\cdot PSW}$ and $\hat{P}_{1\cdot NN10}$ under model A3. The two solid lines indicate the population value P. We can see that overall all point estimators are strongly linearly related. This relationship is, however, less strong when weighting and imputation are compared (figure 5.4 and 5.5). The two scatterplots indicate that under comparable conditions both imputation and weighting can be expected to give very similar results in most cases. However, as can be seen from some outlying values both estimators can produce quite different estimates depending on the sample. Comparing the two imputation methods with each other (figure 5.6) we observe a very strong linear relationship indicating that the two imputation methods lead to very similar point estimates under comparable conditions.

The relationship between the three estimators is further examined using the correlation coefficient. The results are given in table 5.14. The correlation between the weighted and the imputed estimators is positive and close to 1 (between 0.89 and 0.96). This indicates a high linear relationship between the estimators obtained using imputation and weighting. The correlation coefficient between the imputed estimators $\hat{P}_{\cdot HDI10}$ and $\hat{P}_{\cdot NN10}$ is between 0.95 and 0.97 and therefore slightly larger than the correlation between weighting and imputation.

Model	Correlation of $\hat{P}_{\cdot PSW}$ and $\hat{P}_{\cdot HDI10}$	Correlation of $\hat{P}_{\cdot PSW}$ and $\hat{P}_{\cdot NN10}$	Correlation of $\hat{P}_{\cdot NN10}$ and $\hat{P}_{\cdot HDI10}$
A3	0.89	0.90	0.97
A1	0.95	0.94	0.97
A0	0.95	0.95	0.97
DM1	0.96	0.95	0.95
DM2	0.96	0.96	0.96
DM3	0.95	0.94	0.96
DM4	0.95	0.95	0.97

Table 5.14: Correlation coefficients between the estimators based on propensity score weighting (PSW), hot deck imputation within classes without replacement (HDIwor10) and nearest neighbour imputation (NN10).

5.2.5 Application of Predictive Mean Matching Imputation and Propensity Score Weighting to LFS Data

In the following nearest neighbour imputation (NN10), hot deck imputation within classes (HDIwr10 and HDIwor10) and propensity score weighting are applied to LFS data and corresponding estimates for the two proportions of interest, $\hat{P}_1.$ and $\hat{P}_2.$, are presented. In addition, different imputation and propensity score models are used to analyse the effects of various model specifications on estimates of low pay. The analysis is based on two LFS quarters. The quarter June-August 1999 (JA99) refers to a period soon after the introduction of the NMW, however, it is only based on approximately 25% respondents. We therefore also take into account the quarter March-May 2000 (MM00), which is based on approximately 43% respondents and is assumed to be less affected by recent changes in the pay distribution due to the introduction of the NMW in April 1999. We therefore expect to obtain more reliable estimates from the MM00 quarter.

The results for the different estimates are given in table 5.15 and 5.16 based on the quarters June-August 1999 and March-May 2000. For comparison estimates based on the derived variable, without any adjustment for nonresponse bias, are given suggesting a considerable overestimation of the proportion of low paid employees highlighting the need for appropriate adjustment. For

both datasets propensity score weighting gives lower estimates of the proportion of low paid employees under all models considered here in comparison to the imputation methods. For the quarter June-August 1999 the estimates based on the different imputation methods are close to each other with slightly lower estimates for the NN10 method. For this quarter the estimates for weighting and imputation seem to differ, at least slightly, with lower estimates for $\hat{P}_1$. under the weighting method. The slightly lower estimates for $\hat{P}_1$. under propensity score weighting might be explained by the greater negative bias of this method found in the simulation study (table 5.12). The models DM1, DM2 and DM3 produce larger estimates for all methods, which coincides with the strong positive bias for all methods under these models observed in the simulation study (table 5.12). For the quarter March-May 2000, however, where the response rate is higher than in the quarter June-August 1999, the estimates under imputation and weighting are very similar (table 5.16). Taking into account that the quarter MM00 is expected to give more reliable results we conclude that both imputation and weighting seem to produce very similar estimates. Similar results are found in Manning and Dickens (2002), also indicating that propensity score weighting and imputation lead to similar estimates under the assumption of MAR.

In addition, table 5.15 shows the effect of different specifications of the underlying models. Within each method we see that the results for similar models, such as model A3, A1, A0 and DM4, are reasonably close. However, the results can vary considerably for models that depart from these models, for example if they include much less variables or do not include the derived variable X, such as models DM1, DM2 and DM3. We conclude that for both imputation and weighting different models can produce different results for estimates on low pay. There is an indication that both imputation and weighting are sensitive to different model specifications. Careful modelling of either the nonresponse or the prediction of hourly earnings is therefore important.

Method	Model	$\hat{P}_1$. in % (weighted, 18+)	$\hat{P}_2$. in % (weighted, 18+)
Derived Variable	-	7.13	20.51
PSW	A3	0.96	34.52
	A1	1.08	38.41
	A0	1.08	38.40
	DM1	1.33	46.14
	DM2	1.19	41.23
	DM3	1.22	41.61
	DM4	1.07	39.11
NN10	Imputation model as in table 2.1	1.34	32.57
	A3	1.32	32.62
	A1	1.44	32.79
	A0	1.50	33.04
	DM1	1.61	45.07
	DM2	1.70	37.55
	DM3	1.65	37.31
	DM4	1.44	33.32
HDIwr10	Imputation model as in table 2.1	1.50	32.06
	A3	1.47	32.10
	A1	1.41	32.93
	A0	1.54	33.33
	DM4	1.45	33.46
HDIwor10	Imputation model as in table 2.1	1.55	32.04
	A3	1.44	32.11
	A1	1.41	32.91
	A0	1.50	33.20
	DM4	1.45	33.43

Table 5.15: Estimates for $\hat{P}_1$. and $\hat{P}_2$. (weighted) for age group 18+ using different propensity score and imputation models, June-August 1999.

Method	Model	$\hat{P}_1$. in % (weighted, 18+)	$\hat{P}_2$. in % (weighted, 18+)
PSW	A3	0.54	27.10
NN10	A3	0.55	26.61
HDIwr10	A3	0.58	27.03
HDIwor10	A3	0.57	26.01

Table 5.16: Estimates for $\hat{P}_1$. and $\hat{P}_2$. (weighted) for age group 18+ using variables employed in model A3, March-May 2000.

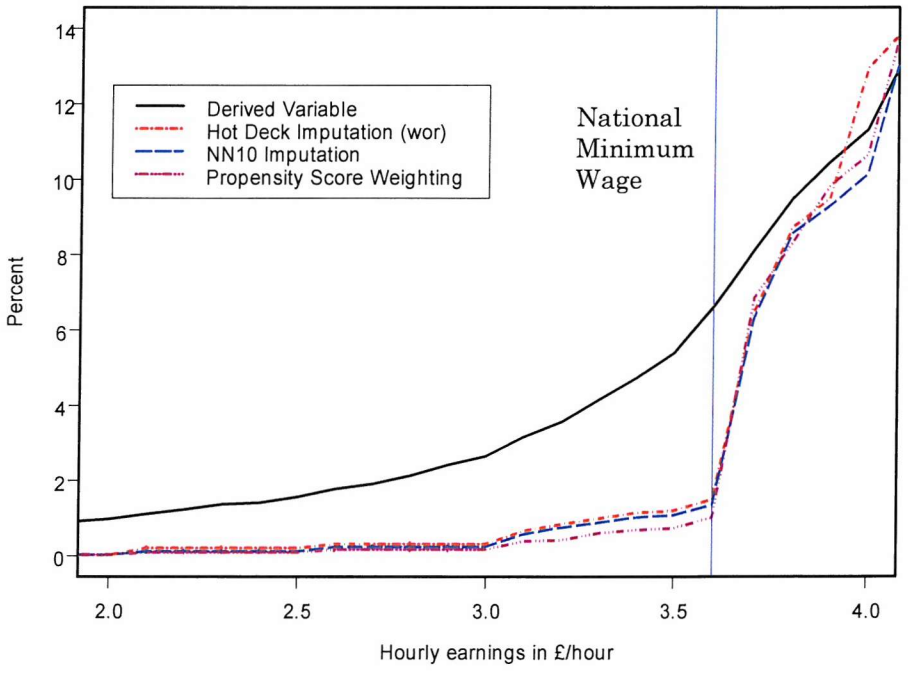

Figure 5.7: Distribution of hourly earnings from £2 to £4 for age group 18+ (weighted), for LFS quarter June-August 1999, based on the derived variable, HDIwor10, NN10 and propensity score weighting.

Figure 5.7 shows estimated distributions of hourly earnings based on the derived variable, hot deck imputation within classes (HDIwor10), NN10 and propensity score weighting. Again we can see that both imputation and weighting produce very similar results with slightly lower estimates

under the weighting approach for the bottom end of the distribution. In all correction methods using either weighting or imputation, we observe a 'kink' in the estimated distributions indicating the level of the NMW as well as other step effects. The estimated distribution based on the derived variable ignoring measurement error suggests in comparison considerable upward bias and stresses the need for adjustment methods.

5.3 Conclusion

This chapter considered a broader class of adjustment methods to compensate for nonresponse bias and to correct for measurement error in an earnings variable. Various forms of predictive mean matching imputation were analysed and compared to propensity score weighting. We found a small positive bias under hot deck imputation within classes which could be caused by the specification and the width of the classes. Different forms of nearest neighbour imputation where therefore analysed, which showed a nonsignificant bias under correct model specifications. The empirical investigation showed that repeated imputation leads to a considerable gain in efficiency compared to single imputation, for example when comparing NN1 and NN10. The gain in efficiency is greater the greater M. The use of a penalty function also reduces the variance of the estimator, however, not as much as the use of repeated imputation. No obvious differences in the performances between random and deterministic forms of nearest neighbour imputation were found. Comparing nearest neighbour imputation with propensity score weighting we found that both imputation and weighting lead to approximately unbiased estimates under certain assumptions such as the assumption of MAR and correct model specifications, however, with some gains in efficiency for the imputation method. In addition, there is an indication that nearest neighbour imputation is more robust against misspecification of underlying assumptions such as misspecification of the imputation model involved and departure from the MAR assumption, in comparison to hot deck imputation within classes and propensity score weighting. Applying the nearest neighbour imputation and propensity score weighting to LFS data we can see that under comparable conditions the two methods produce very similar results. Overall, nearest neighbour imputation based on repeated imputations is recommended for practical use.

Chapter 6

Modelling the Measurement Error and Alternative Estimation Methods

The imputation and weighting methods considered in the previous chapters are based on the assumption of ignorable nonresponse (MAR). It is of interest, however, to explore alternative assumptions, reflecting nonignorable nonresponse taking into account the measurement error structure of the data, and to investigate possible imputation methods based on these assumptions. In this chapter the aim is to investigate alternative imputation methods that are based on an assumption in a measurement error framework, as an alternative to the MAR assumption. Before specifying this assumption it is necessary to investigate the measurement error structure of the derived variable in the LFS.

In general, there are two ways of analysing the nature of measurement error. It is possible to either use validation data giving the true values based on an external or internal source or to use repeated measurements, where all measurements are subject to measurement error (for general approaches to handling measurement error see chapter 1). Preliminary analysis on the earnings variables of interest in the LFS has shown that the derived hourly pay variable X is subject to a considerable amount of measurement error and that the direct variable Y gives much more reliable results on hourly earnings (see section 2.1.3; Skinner et al., 2003; Manning and Dickens, 2002). Based on these findings we regard the direct variable as true earnings, which is not affected by measurement

error (assumption 2.1). Thus, we treat the observed values of Y as internal validation data, not obtained, however, on a randomised subsample of the whole sample, but only on specific members in the sample that fulfil certain criteria, i.e. being paid on an hourly basis, which is not representative for the whole population. The direct variable is therefore missing for a considerable part of the sample s. For salaried employees we therefore can only observe an approximation to Y, the derived variable X, which is, however, subject to measurement error.

6.1 Modelling the Measurement Error

Measurement error in variables may be assumed to be of the classical form (see section 1.1.1), i.e. it is assumed that the measurement error is normally distributed with mean zero and constant variance and that the error is uncorrelated with true earnings and with the explanatory variables, such that

$$x_i = y_i + \delta_i \quad \forall i \in s, \tag{6.1}$$

where

$$\delta_i \sim N(0,\sigma_\delta^2),\ \mathrm{cov}_{DM}(y_i,\delta_j) = 0 \text{ and } \mathrm{cov}_{DM}(\mathrm{w}_i,\delta_j) = 0, \tag{6.2}$$

for all units i and j and explanatory variable W, using the notation introduced in section 1.1.1. However, when analysing the nature of measurement error in earnings variables these assumptions are often violated and systematic error occurs, as discussed in section 1.2. We found that in the LFS when analysing the derived and direct variable some of the assumptions in the classical model hold but that some are clearly violated. To what extent the classical assumptions are violated also depends on the choice of the measurement error model. Several measurement error models are discussed. To facilitate our discussion we introduce the definition for the error terms:

1.) $\delta = X - Y$ (6.3)

2.) $d = \ln(X) - \ln(Y)$ (6.4)

For the following analysis the LFS quarter March-May 2000 (MM00) is used. This dataset contains 15,895 employees in the UK, of whom 6,840 report the direct and the derived variable (excluding employees that are not eligible for the NMW such as people under 18 years old).

Before describing the measurement error models some preliminary comments about outliers with respect to the distribution of Y and X in the LFS sample are made. In general, it is advisable to examine the data for extreme cases, e.g. for extreme differences between X and Y or $\ln(X)$ and $\ln(Y)$, since such outliers might have a great influence on survey estimates. An important source of such outliers are gross errors, such as editing and reporting errors (Hampel et al., 1986), and it is clearly desirable that inference is not based on such errors. In the LFS, some evidence of the presence of such gross errors can be found. For example one case was found that reported hourly earnings of £500 per hour based on the direct variable whereas the derived variable indicated a value just under £5 per hour. Investigating the case it seems likely that the true value should be £5 per hour. One problem, however, is the identification of outlying values. There are several ways of defining outliers (see also Duncan and Hill, 1985; Bound et al. 1994). One possibility is to standardise the error $\delta = X - Y$ and define outliers, for which the standardised error is greater than 3 and smaller than -3. This way 14 extreme cases are identified in the LFS quarter referred to as *type 1* outliers. Another possibility is to define outliers based on the standardisation of the error using the logarithmic transformation, i.e. $d = \ln(X) - \ln(Y)$. This way 120 cases are identified as outliers, referred to as *type 2* outliers. Thus, our attention to outliers is to ensure that the analysis is not driven by a few cases. Another reason for deleting outliers from the dataset is that for the extreme cases there is some reason to believe that the direct variable is subject to error against assumption 2.1. The impact of removing or including outliers in the dataset will be discussed throughout the following.

6.1.1 Measurement Error Models

Several possible measurement error models using additive and multiplicative models are presented as well as analysis of violation of classical measurement error assumptions.

6.1.1.1 Additive Model

Consider the classical measurement error model of the form

$$X = Y + \delta, \text{ where } \delta \sim N(0, \sigma_\varepsilon^2). \tag{6.5}$$

In the following we will discuss the validity of the assumptions made in this model. Figure 6.1 shows the scatterplot of the direct variable Y against the error terms δ, where $\delta = X - Y$. We can see that the error terms are centred around zero. There is evidence that the variance of the error terms increases as Y increases, which means that the assumption of constant variance is violated. There is also an indication of an increasing negative error with increasing Y, whereas for small values of Y, for example at round £4 per hour, more cases show a positive error. For some cases this positive error is large. The scatterplot also shows evidence of a truncation effect due to the NMW level as well as step effects in the variable Y. Note that the scatterplot plotting the derived variable against the direct variable is given in figure 2.2.

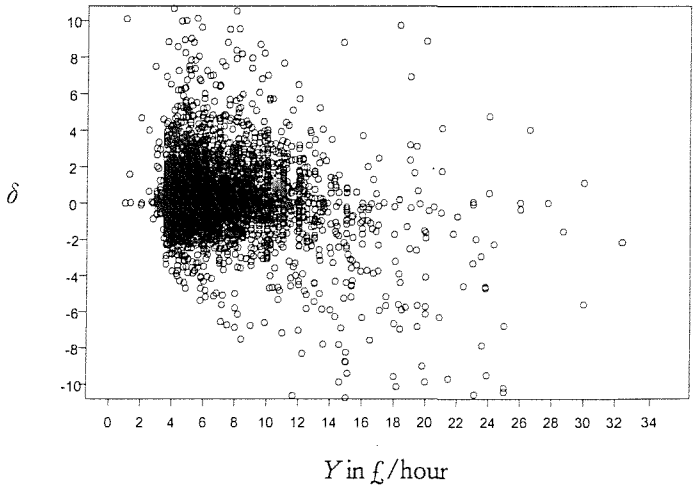

Figure 6.1: Scatterplot of Y against error terms $\delta = X - Y$, not excluding outliers.

Table 6.1 shows the mean and the variance of the error terms as well as the covariance between the error terms and the variable Y, including and excluding type 1 outliers. We can conclude from the results in this table that type 1 outliers have a significant effect on the analysis of the error terms, which can be particularly seen in the difference in variance and covariance of the error terms including or excluding outlying values. Due to the high influence of type 1 outliers it was decided to exclude these cases from further analyses. Using a t-test it is found that the mean of the error terms δ is significantly different from zero on a 99% significance level when excluding type 1 outliers. Including all cases the mean is significantly different from zero using a 95% significance

level. We also analyse the correlation between the error term and the true variable Y. Using Pearson's correlation coefficient, which is valid if both variables are normally distributed, we find a high negative correlation, significantly different from zero. However, the earnings variable Y is not normally distributed and using Spearman's correlation coefficient for a nonparametric test the correlation is almost zero (and not significant). We also find that the error terms δ are not normally distributed, using a Kolmogorov-Smirnov test and by investigating the frequency distribution in figure 6.2, which shows a high spike at zero and longer tails than the normal distribution. Similar results were found in Bound et al. (1994).

Descriptive Statistics for Error Term $\delta = X - Y$		
	Not excluding outliers	Excluding type 1 outliers
Mean	0.168	0.201*
Variance	48.947	3.723
$\text{cov}(\delta, Y)$	-42.652	-1.101
$\text{corr}(\delta, Y)$ (Pearson)	-0.842*	-0.182*
$\text{corr}(\delta, Y)$ (Spearman)	0.007	0.007

Table 6.1: Descriptive statistics for error term δ and effects of type 1 outliers. (A star (*) indicates a significant difference from zero on a 99% significance level.)

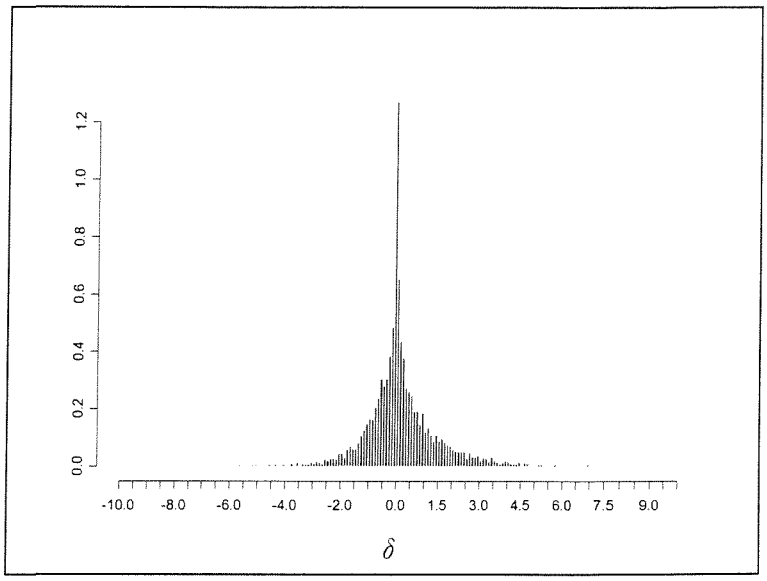

Figure 6.2: Error distribution for $\delta = X - Y$, excluding type 1 outliers.

In addition, the linear regression given in (6.5) is modelled based on respondent cases using the method of least squares, with the derived variable as the dependent and the Y variable as the independent variable, excluding type 1 outliers. The table of coefficients and model diagnostics can be found in appendix A6.1. The intercept of this model is significantly different from zero and the coefficient for the direct variable is not equal to 1. The model is affected by outliers, which can be seen for example in the effect on the R-squared value. Including all cases the R-squared takes on a value of about 0.11, whereas excluding type 1 outliers the R-squared increases to 0.68. Analysing the residual plots of this model we can see that the variance of the residuals decreases as Y increases (figure A6.1.1) and that the residuals are not normally distributed (figure A6.1.2). Also the independence assumption between δ and Y is violated. We conclude that the assumptions made in the additive model (6.5) are clearly violated and that therefore this model is not adequate in describing the nature of the measurement error in the derived variable. Also it was found that type 1 outliers can have a great impact on the performance of this model.

6.1.1.2 Multiplicative Model

Alternatively, we consider the measurement error model to be of the form

$$\ln(X) = \ln(Y) + d \text{, where } d \sim N(0, \sigma_d^2), \tag{6.6}$$

which is equivalent to the multiplicative model of the form $X / Y = \exp(d)$. The advantage is that the error is expressed as the ratio between the erroneous and the true variable, such that the error is given in relative terms instead of absolute values. This is natural since in such earnings variables an error of say £1 is significant for low-earners, whereas for high-earners it might be negligible. Note that the multiplicative model can be expressed as an additive model if we write $X = Y + Y(\exp(d) - 1)$.

Figure 6.3 shows the scatterplot of $\ln(Y)$ against the error term d, where $d = \ln(X) - \ln(Y)$. We can see that the error terms are centred around zero. There is an indication that the variance of the error terms decreases as the direct variable increases. The variance is therefore not constant. The error terms seem to be larger for small values of $\ln(Y)$. Again we observe a truncation effect in the direct variable. Note that the scatterplot of ln(derived variable) against ln(direct variable) is given in figure 2.4.

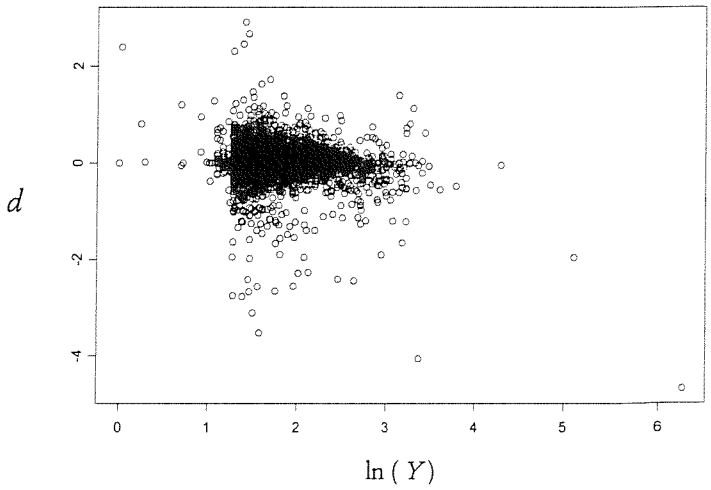

Figure 6.3: Scatterplot of $\ln(Y)$ against the error term $d = \ln(X) - \ln(Y)$, not excluding outliers.

Table 6.2 gives the mean and the variance of the error terms as well as the covariance between the error terms and the variable Y, including and excluding type 2 outliers. Using a t-test there is no evidence that the means of the errors d on either a 99% nor a 95% significance level are significantly different from zero for both cases, using all units and excluding type 2 outliers. Overall, we can see that type 2 outliers do not have a great effect on the results shown in this table. In addition, the correlation between error term d and the true variable $\ln(Y)$ is investigated. As shown in table 6.2 according to Pearson's correlation a negative correlation exists between the error and true earnings, which is significantly different from zero. This is similar to the findings of Bound et al. (1990 and 1994), who also report a negative correlation. However, using Spearman correlation as a nonparametric test we conclude that the correlation is not significantly different from zero. In addition, the association between the error terms and other explanatory variables is analysed. Since most of the explanatory variables are categorical we investigate the differences in means for the error terms within different categories. We found that the means of the error terms for some of the covariates of interest differ significantly in each of the categories, e.g. for the variable Gender. The assumption of no association between the error terms and the covariates is therefore violated.

Descriptive Statistics for Error Term $d = \ln(X) - \ln(Y)$		
	Not excluding outliers	**Excluding type 2 outliers**
Mean	0.008	0.007
Variance	0.093	0.081
$\text{cov}(\delta, Y)$	-0.011	-0.007
$\text{corr}(\delta, Y)$ (Pearson)	-0.096*	-0.064*
$\text{corr}(\delta, Y)$ (Spearman)	0.009	0.009

Table 6.2: Descriptive statistics for error term d and effects of type 2 outliers. (A star (*) indicates a significant difference from zero on a 99% significance level.)

The frequency distribution for errors d are presented in figure 6.4. The distribution is centred around zero and is symmetric, such that under and overreporting cancel out. A similar form of the distribution of the error was found by Bound et al. (1994). However, the error terms do not follow a normal distribution since the error distribution shows a higher frequency for (near-) zero error than normality would imply. For a normal distribution the interquartile range would be 1.33 times the standard deviation. In our analysis the interquartile range (0.195 for error term d) is smaller than the standard deviation (0.305). A similar conclusion was drawn by Bound et al. (1994). Using the Kolmogorov-Smirnov test we conclude that the error terms d are not normally distributed.

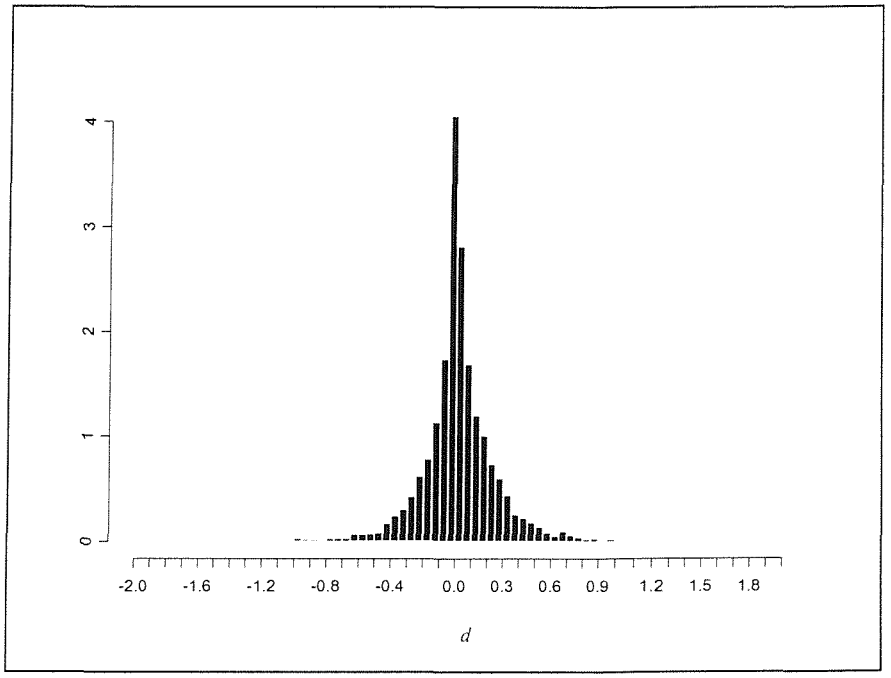

Figure 6.4: Error distribution for $d = \ln(X) - \ln(Y)$, excluding type 2 outliers

In addition, we model the linear regression given in (6.6) using the method of least squares, with the variable $\ln(X)$ as the dependent and the variable $\ln(Y)$ as the independent variable excluding type 2 outliers. The table of coefficients and model diagnostics can be found in appendix A6.2. The coefficient of the intercept is close to zero and the coefficient for $\ln(Y)$ is close to one, which approximates the structure of the proposed multiplicative model. The residual plot 'residual versus fitted values' given in figure A6.2.1 indicates a non-constant variance of the residuals. The

error variance seems to decrease as $\ln(Y)$ increases. Analysing the normal probability plot in figure A6.2.2 we find an indication of non-normality of the residuals, departing from a straight line. Note that the model is not very sensitive to outliers, which can be seen by analysing the effects on the R-squared value or the coefficients and significance of variables when including or excluding outliers. The R-squared value including all cases is 0.55, whereas excluding type 2 outliers the value is 0.61. We conclude that the multiplicative model performs better than the additive model. However, we still observe heteroskedasticity in the error structure. In this case, the following model might be applicable to address the problem of non-constant variance. We have

$$\ln(X) = \ln(Y) + d^{(K)}, \tag{6.7}$$

where $d^{(K)} \sim N(0, \frac{\sigma_d^2}{K})$ and for example $K = 1,2,3,...$, where K is chosen depending on the value of $\ln(Y)$, for example grouping $\ln(Y)$ into K groups with increasing value of $\ln(Y)$.

6.1.1.3 Dependence of Measurement Error on other Covariates

A key issue when modelling the measurement error structure is the extent to which measurement error is systematically related to demographic and economic variables. It is therefore of interest to model the measurement error structure and its dependence on certain explanatory variables in the LFS. We consider a model of the form

$$d_i = \ln(x_i) - \ln(y_i) = g(\ln(y_i)) + \mu_i\beta + \varepsilon_i, \tag{6.8}$$

where the function $g(\ln(y_i))$ allows for non-linearity in the relationship between the error and true earnings and μ_i is a vector of functions of the covariates W for individual i. Under classical assumptions $g(\ln(y_i)) = 0$, $\beta = 0$ and the error term has mean zero and a constant variance. A similar model is discussed in Brownstone and Valetta (1996). Model (6.8) is an extension of the classical measurement error model now taking into account a regression relationship, based on other covariates W and allowing for an intercept and a slope. It is therefore more general than the classical error model. Note that (6.8) describes a model of the form $f(\ln(X) \mid \ln(Y), W, I = 1)$, such that

$$\ln(x_i) = h(\ln(y_i)) + \mu_i\beta + d_i, \tag{6.9}$$

where $h(\ln(y_i))$ is a function of the ln of the direct variable. Several models of the form (6.8) were analysed including or excluding the direct variable Y as an explanatory variable, nonlinear terms, interactions, regarding all cases or removing type 1 or type 2 outliers. Variables found to be significant when trying to explain measurement error in the earnings variable are listed in table 6.3. The table of coefficients as well as residual plots are given in the appendix A6.3. Type 2 outliers are excluded from this model since this was found to improve the fit of the model and the residual plots. The model presented here includes $\ln(Y)$, $\ln^2(Y)$ and some two-way interactions. The residuals $\hat{\varepsilon}_i$ are approximately normally distributed, as shown in the normal probability plot in figure A6.3.2. However, analysing the residual plot of the residuals versus fitted values in figure A6.3.1 we conclude that there is an indication of non-constant variance.

Name of Variable	Abbreviation	Cont./Categ.
Ln of direct hourly pay, $\ln(Y)$	LHR	continuous
Ln of direct hourly pay squared, $\ln^2(Y)$	LHRsq	continuous
Gender	FEMALE	categorical
Additions to basic pay	ADDTBP	categorical
Documentation used to confirm income	USESLP	categorical
Paid less than weekly	LTWK	categorical
Ever work overtime	EVEROT	categorical
Major occupation group	SOCcat	categorical
Last pay same as usual	USGRS	categorical
Married	MARRIED	categorical
Industry sector	INDcat	categorical
Interaction $\ln(Y) \times$ less than weekly	LHR:LTWK	-
Interaction $\ln(Y) \times$ work overtime	LHR:EVEROT	-
Interaction gender× work overtime	FEMALE:EVEROT	-
Interaction additions to pay× less than weekly	ADDTBP:LTWK	-

Table 6.3: Covariates employed in linear regression model modelling the measurement error.

We briefly compare the model in (6.9) given in A6.3 predicting the derived variable with the imputation model predicting the direct variable, given in A2.1. In the model predicting the derived variable less variables are found to be significant than in the model for the direct variable. The model for the derived variable includes 11 variables and two two-way interactions with $\ln(Y)$, whereas the imputation model includes 15 variables with five two-way interaction terms with $\ln(X)$. Both models include variables related to predicting hourly earnings such as major occupation group (SOCcat), Industry section (INDcat) etc. However, the model for the derived variable includes in addition to $\ln(Y)$ mainly variables that are interpreted to be predictors of measurement error, such as additions to basic pay (ADDTBP), ever work overtime (EVEROT), last pay same as usual (USGRS) and documentation used to confirm income (USESLP). The model for the direct variable includes mainly variables related to hourly earnings, such as age (AGE), length of time continuously employed (EMPMONcat), whether a person works part-time (PT) and qualifications (Qcat). This supports the argument that the derived variable is subject to greater measurement error than the direct variable.

Having analysed several measurement error models we conclude that the measurement error in the derived variable does not follow the classical measurement error assumptions as described in chapter 1. In the additive and the multiplicative model it is found that the error terms do not follow a normal distribution and that the assumption of constant variance of the error terms is violated. In addition, the error depends on hourly earnings as well as other covariates. Therefore, correcting for measurement error is not a simple matter and methods such as described in section 1.1.3 might not be adequate. In the following alternative methods correcting for measurement error are explored.

6.2 The Common Measurement Error Assumption

6.2.1 The Common Measurement Error Assumption

An alternative approach to using the assumption of ignorable nonresponse for imputing the missing values is to use the following assumption about the measurement error. The assumption is

referred to as the common measurement error assumption (CME). Based on this assumption the aim is to derive an alternative imputation method for the variable Y.

Assumption 6.1 (CME)

We assume a common measurement error assumption, which says that the variable X is conditionally independent of I given the variables Y and W, i.e.

$$X \perp I \mid (Y,W). \tag{6.10}$$

This is equivalent to $f(X \mid Y,W,I) = f(X \mid Y,W)$

or alternatively $$f(X \mid Y,W,I=0) = f(X \mid Y,W,I=1). \tag{6.11}$$

Proof

$X \perp I \mid (Y,W)$ is defined as $f(X,I \mid Y,W) = f(X \mid Y,W) f(I \mid Y,W)$. Therefore

$$f(X \mid Y,W,I) = \frac{f(X,Y,W,I)}{f(Y,W,I)} = \frac{f(X,Y,W,I)/f(Y,W)}{f(Y,W,I)/f(Y,W)} = \frac{f(X,I \mid Y,W)}{f(I \mid Y,W)} = f(X \mid Y,W)$$

❑

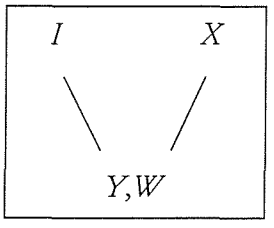

Figure 6.5: Graph of conditional independence under CME assumption.

The graph of conditional independence under the CME assumption is shown in figure 6.5. An advantage of this assumption compared to MAR is that it allows the probability of response to depend on the true level of hourly earnings, Y. The MAR assumption only allows this probability to depend on the measured variable X, which may seem less plausible. The CME assumption means that the variable X has no information about the nonresponse variable I other than what is available in Y and W. This is sometimes referred to as *nondifferential* measurement error and the

variable X is said to be a *surrogate*, since it is conditionally independent of the response given the true variables, here Y and W (Carroll, Ruppert and Stefanski, 1995, p. 16). If we assume a relationship between Y and X such as $X = Y + \delta$ then (6.10) can be written as $\delta \perp I | (Y,W)$, which means that the measurement error is independent of the probability of response given Y and W. Since of course $X \perp I | (Y,W) \Leftrightarrow I \perp X | (Y,W)$, we can also write $f(I | X,Y,W) = f(I | Y,W)$, which means that the probability of response, i.e. the distribution of I, depends on the variables Y and W but not on the derived variable X, conditioning on Y and W. The assumption of CME implies that the measurement error process is the same for employees paid hourly and those who are not. However, also this assumption is a restricting assumption which might not hold in reality. For example, one might assume that a person that is paid by the hour might have a better knowledge about his pay details and in particular hours worked than a salaried person such that the measurement errors for those paid hourly and those who are not are likely to be different. It can be easily seen that both assumptions, the above common measurement error assumption (CME) and the assumption of data missing at random (MAR), can hold at the same time. In fact it holds: The assumptions MAR and CME are equivalent to the fact that the response mechanism does neither depend on Y nor on X, but possibly on W, i.e.

$$I \perp Y | (X,W) \text{ and } X \perp I | (Y,W) \Leftrightarrow I \perp (Y,X) | W. \tag{6.12}$$

A special case is if the data are missing completely at random (MCAR). The implications in (6.12) follow according to the block independence lemma (Whittaker, 1990, p. 33). Also for '$\Leftarrow$' it can be easily seen that if the response mechanism I does neither depend on Y nor on X, possibly conditional on W, i.e. $f(I | Y,X,W) = f(I | W)$, it follows $f(I | Y,W) = f(I | W)$ and $f(I | X,W) = f(I | W)$. Therefore we have $f(I | X,Y,W) = f(I | X,W)$ and $f(I | X,Y,W) = f(I | Y,W)$, which are the definitions of MAR and CME.

As in section 2.2.2 the aim is to generate $Y_{\bar{r}}$ from the covariates X and W for I=0, such that

$$f(Y. | X,W,I = 0) = f(Y | X,W,I = 0), \tag{6.13}$$

which means that the distribution of the imputed values given X and W is the same as the conditional distribution of the true values. We are therefore interested in the estimation of the

distribution $f(Y \mid X, W, I=0)$ using the assumption of CME. First, the performance of the imputation and weighting methods, developed under the MAR assumption, are evaluated under the assumption of CME. In section 6.3 several approaches are described for estimating the distribution of interest, $f(Y \mid X, W, I=0)$, using the assumption of common measurement error.

6.2.2 Performance of Predictive Mean Matching Imputation and Propensity Score Weighting under the Assumption of Common Measurement Error

It is of interest to evaluate the performance of the imputation and weighting methods considered in the previous chapters under the alternative assumption of CME. Nonresponse under the CME assumption reflects misspecification of the nonresponse mechanism that these imputation and weighting methods are theoretically based on. In the following, we investigate various forms of predictive mean matching imputation, i.e. hot deck imputation within classes with and without replacement (HDIwr10 and HDIwor10) and nearest neighbour imputation with 10 imputed values (NN10), as well as propensity score weighting (PSW) under the CME assumption using a simulation study.

a.) Design of the Simulation Study

The generation of the variables for each sample $s^{(a)}$, $a=1,....,A$, is carried out as described in section 3.3.1. The covariates W are obtained by bootstrapping the original sample s. The variables $\ln(X)$ and $\ln(Y)$ are generated by using the predictions of a linear regression model and adding on random errors, as described in equations (3.20) and (3.21). Nonresponse is introduced according to the assumption of CME. To do this the probabilities of response p_i, for all $i \in s^{(a)}$, need to be specified conditionally independently of X given Y and W. Estimates of these probabilities are obtained using a logistic model of the form

$$log(\frac{\hat{p}_i}{1-\hat{p}_i}) = \hat{\varphi}_1 + \hat{\varphi}_2 \ln(y_i) + \sum_{v=1}^{V} \hat{\varphi}_v w_{vi} \quad \text{for all } i = 1, \ldots, n, \tag{6.14}$$

where the estimated coefficients are obtained using the logistic regression model specified in (3.23), including the variable $\ln(X)$ instead of $\ln(Y)$ since $\ln(X)$ is observed for all $i \in s$. A nonresponse mechanism is therefore generated that depends on Y and W but not on X. The imputation model used for hot deck imputation within classes (HDI) and nearest neighbour (NN)

is specified in section 2.2.4. For propensity score weighting (PSW) the model A3 is used as specified in section 3.3.4. The overall response rate using the assumption of CME is approximately 45% and therefore slightly higher than under nonresponse following MAR with approximately 43%.

b.) Performance of Imputation and Weighting Methods under the CME Assumption

In the following the simulation results for PSW, NN10 and HDI under the assumption of CME are investigated. The imputation class boundaries for HDI are defined in the same way as in section 3.3. Table 6.4 shows bias and relative bias for both point estimators $\hat{P}_1.$ and $\hat{P}_2.$ and the variance and mean square error of $\hat{P}_1.$. For all methods the biases for both point estimators are with around 5% to 6% significantly higher than in the case of nonresponse following the MAR assumption. For both point estimators HDI shows the highest relative bias with around 6%. The relative bias under PSW is slightly smaller with around 5.7%. The smallest relative bias with approximately 5% is obtained under NN10, which overall also gives the smallest mean square error. Finding a significant increase in the bias of the estimators under the CME assumption as opposed to the MAR assumption makes it worthwhile investigating adjustment methods that are based on the assumption of CME instead of MAR and therefore correct for the bias found in this analysis.

Method	Bias of $\hat{P}_1.$	Rel. Bias of $\hat{P}_1.$	Bias of $\hat{P}_2.$	Rel. Bias of $\hat{P}_2.$	$V(\hat{P}_1.)$	$MSE(\hat{P}_1.)$
PSW	$3.23*10^{-3}$ $(7.32*10^{-5})^*$	5.72 %	$1.0*10^{-2}$ $(1.40*10^{-4})^*$	5.70 %	$5.35*10^{-6}$	$1.58*10^{-5}$
NN10	$2.9*10^{-3}$ $(8.02*10^{-5})^*$	5.14 %	$9.2*10^{-3}$ $(1.48*10^{-4})^*$	4.96 %	$6.44*10^{-6}$	$1.48*10^{-5}$
HDIwr10	$3.41*10^{-3}$ $(6.99*10^{-5})^*$	6.04 %	$1.5*10^{-2}$ $(1.28*10^{-4})^*$	6.20 %	$4.89*10^{-6}$	$1.65*10^{-5}$
HDIwor10	$3.42*10^{-3}$ $(6.96*10^{-5})^*$	6.05 %	$1.2*10^{-2}$ $(1.28*10^{-4})^*$	6.19 %	$4.85*10^{-6}$	$1.65*10^{-5}$

Table 6.4: Simulation results for propensity score weighting (PSW), nearest neighbour imputation (NN10) and hot deck imputation with and without replacement (HDIwr10 and HDIwor10) under the assumption of common measurement error (CME).

6.3 Alternative Estimation Methods Under the CME Assumption

6.3.1 Density Estimation using Deconvolution

As discussed in section 6.2.1 we are interested in estimating the distribution $f(Y \mid X, W, I = 0)$, which, using Bayes' theorem, is of the form

$$f(Y \mid X, W, I = 0) \propto f(X \mid Y, W, I = 0) f(Y \mid W, I = 0). \tag{6.15}$$

The first part on the right hand side, $f(X \mid Y, W, I = 0)$, can be estimated under the CME assumption using data on the respondents only, i.e. $f(X \mid Y, W, I = 1)$. The second distribution on the right hand side can be obtained as part of the following integral

$$f(X \mid W, I = 0) = \int f(X \mid Y, W) f(Y \mid W, I = 0) dY, \tag{6.16}$$

where $f(X \mid W, I = 0)$ and $f(X \mid Y, W)$ can be estimated using the observed data and the common measurement error assumption. To obtain the distribution $f(Y \mid W, I = 0)$ in the absence of parametric assumptions, deconvolution is necessary as described in Carroll, Ruppert and Stefanski (1995), Stefanski and Carroll (1990) and Carroll and Hall (1988). The distribution $f(Y \mid W, I = 0)$ can be recovered by Fourier inversion if both distributions $f(X \mid W, I = 0)$ and $f(X \mid Y, W)$ are known. However, in our example both distributions are unknown and need to be estimated. This can be done using Kernel density estimation. For estimating such functions the choice of kernel is relatively unimportant. However, for commonly used kernels the estimated density $\hat{f}(Y \mid W, I = 0)$ cannot be deconvolved since the integral in the Fourier inversion is not defined. Stefanski and Carroll (1990) showed that for certain smooth kernels the Fourier inversion is possible and an estimate of the distribution of interest could be obtained. However, convergence rates of the estimator $\hat{f}(Y \mid W, I = 0)$ are extremely slow and imply that it is not possible to estimate $f(Y \mid W, I = 0)$ well (Caroll, Ruppert and Stefanski, 1995; Luo, Stokes and Sager, 1998).

Caroll, Ruppert and Stefanski (1995) discuss the problem of estimating the density of Y using deconvolution techniques, when only the variable X containing measurement error is observed and a model of the form

$$\ln(X) = \ln(Y) + d \tag{6.17}$$

is assumed. They found that if the error distribution in (6.17) is normal, $f(Y)$ cannot be estimated well. If the error distribution is more peaked than the normal distribution then the deconvoluting kernel density estimator has a slightly better performance. Caroll, Ruppert and Stefanski (1995) show that the smoothness of the error density determines how well $f(Y)$ can be estimated, which is a disconcerting nonrobustness result. The slow rate of convergence of $\hat{f}(Y)$ applies generally to such deconvolution problems and is not a specific problem of the kernel density estimator used. In addition, estimating quantiles of the distribution of Y without making parametric assumptions indicates similar problems. Due to the difficulties encountered in such density estimation problems, the approach of deconvolution was not investigated further.

6.3.2 Parametric Methods

An alternative to such a nonparametric approach is to use parametric methods. Several methods exist, which make distributional assumptions, for example about the measurement error. The SIMEX method (Simulation-Extrapolation) requires that the form and the distribution of the measurement error are known, such as normality, zero mean and constant variance of the error terms. It also makes the assumption of an additive measurement error model. An estimate of the error variance is needed (Caroll, Ruppert and Stefanski, 1995; Luo, Stokes and Sager, 1998). Stefanski and Bay (1996) propose an estimator of a cumulative distribution function in the presence of measurement error based upon the SIMEX method. Simulation results in Luo, Stokes and Sager (1998) suggest that the SIMEX estimators are sensitive to misspecification of the underlying error assumptions. Because of the restricting parametric specifications required these parametric methods have not been used here. Buonaccorsi (1990) considers a wider class of measurement error models relating true and observed values other than the classical additive models, making assumptions of normality. Maximum likelihood estimators and their properties are developed for the models proposed. However, for this approach true values obtained from a random subsample are required. Luo, Stokes and Sager (1998) propose estimators of a cumulative

distribution function in the presence of a two-phase sample, also called calibration sample, where in a second phase sample, based on a sampling design, true measurements of the variable of interest are obtained. Nusser et al. (1996) develop an estimator of a cumulative distribution function by first using a transformation to normality and then the normal components of variance model. The problem of estimating the density of Y is closely related to estimating the regression function $E(Y \mid X,W)$ when only X, $X = Y + \varepsilon$, and W are observed for all sample units, i.e. obtaining the predicted values for the variable Y conditioning on the variables X and W. This technique is referred to as *regression calibration* (Caroll, Ruppert and Stefanski, 1995; Brown, 1993) and requires for example a validation sample on a subset observing the true values Y or replication data. However, such methods are not applicable here since no random calibration sample including the true variable Y is available.

6.3.3 Measurement Error Approach using Discretized Variables

Since the aim is to estimate proportions of low paid it is possible to use adjustment methods developed for misclassified categorical data to correct for measurement error in the derived variable. For this purpose the variables X and Y are discretized, for example by rounding to the nearest 5p, with categories k, $k = 1,...,J$, and j, $j = 1,...,J$, respectively. The idea is to multiply a vector of observed proportions based on the variable X by an adjustment matrix which leads to estimates of the true proportions based on the variable Y. In the following a method is presented where the adjustment matrix is the inverse of the misclassification matrix containing misclassification probabilities, introduced in section 1.1.2. An advantage of the method is that no assumption about the distribution of Y is made. Such matrix methods have been described in detail by Selén (1986) and Kuha and Skinner (1997). The values of the true variable Y are treated as fixed and the values of the variable subject to misclassification X are treated as random obtained from Y by a random process based on the misclassification probabilities

$$\mathrm{P}(X = k \mid Y = j) = m_{kj}. \tag{6.18}$$

Let $\boldsymbol{P}_Y = (P_Y(Y = 1),...,P_Y(Y = J))'$ be the column vector of proportions of units belonging to the categories of variable Y in the population and $\boldsymbol{M}_{XY} = [m_{kj}]$ the misclassification matrix of probabilities that a unit from category j of variable Y is classified as category k of variable X. Note

that $\sum_k m_{kj} = 1$ and that the parameters of interest P_1 and P_2 as defined in section 3.3 can be obtained from $\boldsymbol{P}_Y$ by summing over appropriate elements $P_Y(.)$. The vector $\boldsymbol{P}_X$ denotes the corresponding vector based on the variable X. We have

$$\boldsymbol{P}_X = \boldsymbol{M}_{XY}\boldsymbol{P}_Y \tag{6.19}$$

where the kth row can be written as

$$P_X(X=k) = \sum_{j=1}^{J} \mathrm{P}(X=k \mid Y=j) P_Y(Y=j). \tag{6.20}$$

We assume that $\boldsymbol{M}_{XY}^{-1}$ exists since a misclassification matrix is expected to have rather large elements on the main diagonal, otherwise the misclassification problem would be of limited value (Selén, 1996, p. 77). We obtain

$$\boldsymbol{P}_Y = \boldsymbol{M}_{XY}^{-1}\boldsymbol{P}_X. \tag{6.21}$$

Let $\hat{\boldsymbol{P}}_X$ denote an estimator of $\boldsymbol{P}_X$ based on the variable X computed from the sample s such that $\hat{\boldsymbol{P}}_X = (\hat{P}_X(X=1),...,\hat{P}_X(X=J))'$, where

$$\hat{P}_X(X=k) = \frac{1}{n}\sum_{i\in s} I(x_i = k). \tag{6.22}$$

For simplicity the survey weights w_i are not taken into account. Note that for units i in sample s only the misclassified categories based on the variable X are observed. The relationship between $\hat{\boldsymbol{P}}_X$ and $\boldsymbol{P}_Y$ can be described by

$$E_{DM}(\hat{\boldsymbol{P}}_X) = \boldsymbol{M}_{XY}\boldsymbol{P}_Y, \tag{6.23}$$

where the expectation is with respect to the misclassification model M and the sampling design D (Kuha and Skinner, 1997). Let $\hat{\boldsymbol{M}}_{XY}$ be a consistent estimate of $\boldsymbol{M}_{XY}$, based on information from external or internal validation data, obtained on a random subsample of the sample s. Assuming that $\hat{\boldsymbol{M}}_{XY}^{-1}$ exists, which is the case for large samples s and sensible misclassification problems (Selén, 1996, p. 77), we obtain a consistent estimate of $\boldsymbol{P}_Y$ by

$$\hat{\boldsymbol{P}}_Y^M = \hat{\boldsymbol{M}}_{XY}^{-1}\hat{\boldsymbol{P}}_X. \tag{6.24}$$

The idea of using the misclassification matrix, which can be simplified under the common measurement error assumption, to obtain estimates of proportions of low paid employees has been employed by Manning and Dickens (2002). They present a method, which provides an estimate of an upper bound of the proportion of employees being paid *at* the NMW, for example being paid between £3.59 and £3.61 for 22+ years old (see chapter 2 for the level of the NMW). This proportion of interest is denoted $\hat{P}_3$.. It is difficult, however, to provide estimates of other parts of the hourly pay distribution based on this method. The method is therefore not attractive as a general technique for estimating hourly pay distributions in the application considered here. However, an advantage is that it is a nonparametric method making no assumptions about the underlying distributions. In the following, this estimation method is briefly discussed.

The variables Y and X are discretised rounding values to the nearest 5p, such that we have, using equation (6.20),

$$P_X(X=k\,|\,W,I=0)=\sum_{j=1}^{J}\mathrm{P}(X=k\,|\,Y=j,W,I=1)P_Y(Y=j\,|\,W,I=0), \qquad (6.25)$$

where k and j are categories of the discretized variables X and Y. Since the proportion of employees earning at the NMW, $\hat{P}_3$., is of particular interest, Dickens and Manning (2002) concentrate on the category where Y indicates that an employee is paid at this rate, denoted $Y=nmw$, and write (6.25) as

$$P_X(X=nmw|\,W,I=0)\geq \mathrm{P}(X=nmw|\,Y=nmw,W,I=1)P_Y(Y=nmw|\,W,I=0), \qquad (6.26)$$

where the summation is replaced by a lower bound since only the category of Y, where $Y=nmw$, is taken into account. This leads to an upper bound for the distribution of interest giving

$$P_Y(Y=nmw|\,W,I=0)\leq \frac{P_X(X=nmw|\,W,I=0)}{\mathrm{P}(X=nmw|\,Y=nmw,W,I=1)}. \qquad (6.27)$$

Both probabilities on the right hand side can be estimated using observed data only. Formula (6.27) gives an upper bound of the proportion of employees being paid at the NMW among the nonrespondents. It is obviously required that the estimate of the denominator is unequal to zero, which means that there must be a group of employees that report on both variables, X and Y, that they are paid at the NMW. Note that the estimation technique using discretized variables cannot

be easily extended to estimate other parts of the distribution of interest due to very small sample sizes within some of the categories required. In particular, the denominator of (6.27) requires a group of employees that on both variables, X and Y, report the same value, i.e. within category k there must be a group of employees who are not misclassified and therefore do not report a measurement error. This method may therefore not be suitable for obtaining general estimates of the distribution of interest.

To obtain an estimate of the marginal distribution, i.e. the proportion of employees paid at the NMW, $\hat{P}_3.$, taking into account respondents and nonrespondents we need to obtain estimates of the following distributions

$$P_Y(Y = nmw \mid W) = P_Y(Y = nmw \mid W, I = 1) * P(I = 1 \mid W)$$

$$+P_Y(Y = nmw \mid W, I = 0) * P(I = 0 \mid W). \qquad (6.28)$$

Since the second term on the right hand side of (6.28) can only be estimated using the upper bound in (6.27) we can only obtain an upper bound for $\hat{P}_3.$. The method is applied to several quarters of the LFS and compared to estimates obtained using propensity score weighting (PSW), nearest neighbour imputation (NN10) and hot deck imputation within classes (HDIwor10) as valid under the MAR assumption. If covariates W are included in the estimation process logistic regression models are used to estimate the probabilities involved. The covariates W used are gender, age, occupation and industry section. The results are given in table 6.5.

For the quarter MM00 the method using discretized variables gives an upper bound of the proportion of interest. All estimates based on the imputation and weighting methods are below this threshold, which means that the method using discretized variables does not gain much. Note, that the estimates based on weighting and imputation are quite similar. For the LFS quarter JA99 we found that the method using discretized variables gives an estimate that is below the estimates obtained from using imputation and weighting, suggesting that the imputation and weighting methods overestimate the true value. Note that the particular quarter JA99 contains only 25% respondents. The estimates obtained are therefore subject to greater variability. Quarters from MM00 onwards contain approximately 43% respondents.

Method	$\hat{P}_3$. (JA99)	$\hat{P}_3$. (MM00)
ME-approach using discretized variables (upper bound), without covariates W	3.21%	2.81%
ME-approach using discretized variables (upper bound), with covariates W	3.22%	2.82%
PSW under MAR	5.14%	2.49%
NN10 under MAR	4.61%	2.56%
HDIwor10 under MAR	4.57%	2.55%

Table 6.5: Estimates of the percentages of employees paid at the NMW for 22+ (unweighted) for the LFS quarters JA99 and MM00, 'at NMW' is defined as $£3.59 \leq Y \leq £3.61$.

6.3.4 Imputation using a Weighted Bootstrap Method Based on the Assumption of CME

Another possibility for estimating the distribution of interest is to use Bayes' theorem and to express the distribution $f(Y \mid X, W, I = 0)$ as the posterior distribution. Since the posterior distribution is often difficult to sample from, such an approach requires additional methods that allow sampling from this distribution. Methods such as rejection sampling, importance sampling, weighted bootstrap and sampling-importance resampling are commonly used to approach this problem (Tanner, 1996; Carroll, Ruppert and Stefanski, 1995; Gilks, Richardson and Spiegelhalter, 1996). For more complex situations iterative methods using a Markov chain, such as Gibbs sampling or data augmentation, are required. In section 6.3.5 the use of data augmentation under the CME assumption is derived. In the following a method is presented that uses the weighted bootstrap to allow drawing values from the posterior distribution. For simplicity, in this section we use the notation Y and X. In the actual application to LFS data and in the simulation study, however, functions of Y and X, such as $\ln(Y)$ and $\ln(X)$ are used.

Result 6.1:

Under the CME assumption the distribution $f(Y \mid X, W, I = 0)$ can be decomposed as

$$f(Y \mid X, W, I = 0) \propto f(Y \mid X, W, I = 1)\lambda(Y \mid W), \tag{6.29}$$

where $\lambda(Y \mid W) = \dfrac{f(Y \mid W, I = 0)}{f(Y \mid W, I = 1)}$.

Proof

$$f(Y \mid X, W, I = 0) \propto f(X \mid Y, W, I = 0)\ f(Y \mid W, I = 0)$$

$$= f(X \mid Y, W, I = 1)\ f(Y \mid W, I = 1)\frac{f(Y \mid W, I = 0)}{f(Y \mid W, I = 1)}$$

$$\propto f(Y \mid X, W, I = 1)\frac{f(Y \mid W, I = 0)}{f(Y \mid W, I = 1)},$$

using Bayes theorem and the CME assumption. ❑

Result 6.1 proposes an imputation method for the missing values of Y under the CME assumption. It allows drawing from the unobserved distribution $f(Y \mid X, W, I = 0)$ by drawing values from the observed distribution $f(Y \mid X, W, I = 1)$ and taking into account the additional factor $\lambda(Y \mid W)$. This imputation method extends the previously discussed predictive mean matching imputation methods, which only take into account the observed distribution $f(Y \mid X, W, I = 1)$. There are several ways of taking into account the additional factor $\lambda(Y \mid W)$ in (6.29), for example using rejection sampling or the weighted bootstrap method (Carroll, Ruppert and Stefanksi, 1995), which will be discussed in the following.

The regression model for $f(Y \mid X, W, I = 1)$ is assumed to be of the form

$$y_i = \eta_i\beta + \varepsilon_i, \tag{6.30}$$

where η_i is a row-vector of functions of the derived variable and other covariates and estimates for the parameters of the model can be obtained using the least squares method based on the respondents. A value $\hat{y}_j^*$ for $j \in \bar{r}$ is generated from the estimated distribution, i.e. $\hat{y}_j^* \sim \hat{f}(y \mid x_j, \mathrm{w}_j, \varsigma)$, where ς is the vector of parameters estimated based on respondent data, for example by obtaining the predicted value $\eta_j \hat{\beta}$ and adding on a random residual $\hat{\varepsilon}_j \sim N(0, \hat{\sigma}_\varepsilon^2)$. Alternatively, the value $\hat{y}_j^*$ can be selected as the value of the nearest neighbour of $j \in \bar{r}$ or by randomly drawing a value $\hat{y}_j^*$ within classes defined for $j \in \bar{r}$ as it is done in nearest neighbour imputation and hot deck imputation within classes respectively. Suppose that the function $\lambda(Y \mid W)$ can be estimated and takes on values between zero and one, then it is possible to obtain imputed values for Y by drawing a value $\hat{y}_j^*$ from the estimated distribution $\hat{f}(Y \mid X, W, I = 1)$ and in addition applying rejection sampling by $\hat{\lambda}(Y \mid W)$. The value $\hat{y}_j^*$ is accepted with probability $\hat{\lambda}(\hat{y}_j^* \mid \mathrm{w}_j)$. If not rejected the value $\hat{y}_j^*$ is the imputed value for Y for the nonrespondent j, i.e. $\hat{y}_j^* = \hat{y}_j$. The procedure is therefore as follows:

1. generate a value $\hat{y}_j^*$ from $\hat{f}(Y \mid X, W, I = 1)$ as under the assumption of ignorable nonresponse
2. generate δ from a Unif(0,1), a uniform distribution on the interval (0,1),
3. if $\delta \leq \hat{\lambda}(\hat{y}_j^* \mid \mathrm{w}_j) \Rightarrow$ accept $\hat{y}_j^*$ as the imputed value for person $j \in \bar{r}$, else go back to 1.

This method, however, requires that $\lambda(Y \mid W)$ lies between zero and one. In our application we found that in general $\hat{\lambda}(Y \mid W)$, estimated under certain distributional assumptions, is not of this form. If $\lambda(Y \mid W)$ exceeds the value one but $\lambda(Y \mid W)$ is bounded, i.e. there is a finite known constant $M > 0$, such that $\lambda(Y \mid W) \leq M$, then we can accept the imputed value with probability $p(Y \mid W) = \lambda(Y \mid W)/M$. However, if M is large and for most values of Y it is $\lambda(Y \mid W) \ll M$, the probabilities $p(Y \mid W)$ are small for most values and it might take a long time until acceptance. If, however, $\lambda(Y \mid W)$ is not bounded, we may draw Q values for person j from $\hat{f}(Y \mid X, W, I = 1)$ denoted by $\hat{y}_{j1}^*, \ldots, \hat{y}_{jQ}^*$ and obtain the values $\hat{\lambda}(\hat{y}_{j1}^* \mid \mathrm{w}_j,), \ldots, \hat{\lambda}(\hat{y}_{jQ}^* \mid \mathrm{w}_j)$, assuming that $\lambda(Y \mid W)$ can be estimated by $\hat{\lambda}(Y \mid W)$. One value denoted $\hat{y}_j$, the finally imputed value, is sampled out of the Q possible values $\hat{y}_{j1}^*, \ldots, \hat{y}_{jQ}^*$ with probabilities

$$\hat{p}_{jq}^* = \hat{\lambda}(\hat{y}_{jq}^* \mid \mathrm{w}_j) / \sum_{q=1}^{Q} \hat{\lambda}(\hat{y}_{jq}^* \mid \mathrm{w}_j) \text{ for all } q = 1, \ldots, Q. \tag{6.31}$$

This method is referred to as the *weighted bootstrap* method (Tanner, 1996, p. 54; Carroll, Ruppert and Stefanski, 1995, p. 170). Rubin (1987) uses this idea as part of the sampling-importance resampling method. The idea is that if Q increases the approximation to the distribution of the nonrespondents improves and the distribution of $\hat{y}_j$ approaches the distribution $f(Y \mid X, W, I = 0)$. The method described in (6.29) is in the following referred to as the *weighted bootstrap imputation* method.

Based on the *weighted bootstrap imputation* method the estimator $\hat{P}.$ can be written as

$$\hat{P}. = \frac{1}{n} \sum_{j \in s} z._j , \tag{6.32}$$

where $z._j = I(y._j < t)$, $y._j = \begin{cases} y_j \text{ for } j \in r \\ \hat{y}_j \text{ for } j \in \bar{r} \end{cases}$ and $\hat{y}_j \in \{\hat{y}^*_{j1}, ..., \hat{y}^*_{jQ}\}$ with probability $\hat{p}^*_{jq}$, for $q = 1, ..., Q$. An alternative way of defining the weighted bootstrap method is to use all selected donor values $\hat{y}^*_{j1}, ..., \hat{y}^*_{jQ}$ for imputation for $j \in \bar{r}$, however weighted by the probabilities of selection $\hat{p}^*_{j1}, ..., \hat{p}^*_{jQ}$ as defined in (6.31). We obtain an estimator of the form

$$\hat{P}. = \frac{1}{n} \left[\sum_{j \in r} z_j + \sum_{j \in \bar{r}} \sum_{q=1}^{Q} \hat{p}^*_{jq} \hat{z}_{jq} \right], \tag{6.33}$$

(note that $\sum_{q=1}^{Q} \hat{p}^*_{jq} = 1$), which effectively means using weighted repeated imputation or fractional imputation, where the weights are derived using the weighted bootstrap method. We will refer to this method as *weighted bootstrap fractional imputation*. The advantage of this weighting method is a potential reduction in variance. The results of the simulation study analysing the weighted bootstrap imputation method suggest a gain in efficiency by using this weighted estimator.

We now turn to the problem of estimating $\lambda(Y \mid W)$. Since the function $\lambda(Y \mid W)$ is generally not known, an estimate for $\lambda(Y \mid W)$ is required when estimating the target distribution $f(Y \mid X, W, I = 0)$. The expression for $\lambda(Y \mid W)$ requires information on the distributions $f(Y \mid W, I = 0)$ and $f(Y \mid W, I = 1)$. The latter one can be estimated using the observed data,

however, $f(Y \mid W, I=0)$ is unknown and certain assumptions are necessary for estimating this distribution. To facilitate our discussion we will introduce some notation for the distributions of interest. To proceed, we assume the following linear relationships.

Assumption A:

Distribution	Mean	Variance
$f(Y \mid W, I=0)$	$E(Y \mid W, I=0) = \mu\Phi_0$	$\text{var}(Y \mid W, I=0) = \sigma_0^2$
$f(Y \mid W, I=1)$	$E(Y \mid W, I=1) = \mu\Phi_1$	$\text{var}(Y \mid W, I=1) = \sigma_1^2$
$f(X \mid W, I=0)$	$E(X \mid W, I=0) = \mu\delta_0$	$\text{var}(X \mid W, I=0) = \varrho_0^2$
$f(X \mid W, I=1)$	$E(X \mid W, I=1) = \mu\delta_1$	$\text{var}(X \mid W, I=1) = \varrho_1^2$
$f(X \mid Y, W, I=0) =$ $f(X \mid Y, W, I=1)$ (because of CME)	$E(X \mid Y, W, I=0) =$ $E(X \mid Y, W, I=1) = Y\beta_Y + \mu\beta_W$	$\text{var}(X \mid Y, W, I=0) =$ $\text{var}(X \mid Y, W, I=1) = v^2$
$f(Y \mid X, W, I=0)$	$E(Y \mid X, W, I=0) = X\alpha_{X0} + \mu\alpha_{W0}$	$\text{var}(Y \mid X, W, I=0) = \pi_0^2$
$f(Y \mid X, W, I=1)$	$E(Y \mid X, W, I=1) = X\alpha_{X1} + \mu\alpha_{W1}$	$\text{var}(Y \mid X, W, I=1) = \pi_1^2$

where μ is a row-vector of functions of the covariates W.

Note that in general we cannot expect that both $E(Y \mid X, W, I=0)$ and $E(Y \mid X, W, I=1)$ are linear functions of X and μ. However, if either we assume normality of the distributions or if we express the non-linear function as approximately linear using Taylor linearisation methods we may proceed with the assumption of linearity. The parameters Φ_1, σ_1^2, δ_0, ϱ_0^2, δ_1, ϱ_1^2, β_Y, β_W, v^2, α_{X1}, α_{W1} and π_1^2 can be estimated from the data using least squares regression, which does not require the assumption of normality of residuals (Draper and Smith, 1998, p. 136). Estimates of the parameters Φ_0 and σ_0^2 can be obtained using estimates from the distributions of $f(X \mid W, I=0)$ and $f(X \mid Y, W, I=0)$ and the assumption of CME. Expressions for Φ_0 and σ_0^2 are obtained as follows:

$$\Phi_0 = \frac{\delta_0 - \beta_W}{\beta_Y} \text{ and} \tag{6.34}$$

$$\sigma_0^2 = \frac{\varrho_0^2 - v^2}{\beta_Y^2} . \tag{6.35}$$

Proof of (6.34) and (6.35)

To obtain (6.34):

$$\begin{aligned}\mu\delta_0 &= E(X \mid W, I=0) = E[E(X \mid Y, W, I=0) \mid W, I=0] \\ &= E[E(X \mid Y, W, I=1) \mid W, I=0] = E[Y\beta_Y + \mu\beta_W \mid W, I=0] \\ &= E[Y \mid W, I=0]\beta_Y + \mu\beta_W = \mu\Phi_0\beta_Y + \mu\beta_W\end{aligned}$$

$$\Leftrightarrow \quad \delta_0 = \Phi_0\beta_Y + \beta_W$$

$$\Leftrightarrow \quad \Phi_0 = \frac{\delta_0 - \beta_W}{\beta_Y} \text{ , assuming } \beta_Y \neq 0 .$$

To obtain (6.35):

$$\begin{aligned}\varrho_0^2 &= \operatorname{var}(X \mid W, I=0) \\ &= \operatorname{var}[E(X \mid Y, W, I=0) \mid W, I=0] + E[\operatorname{var}(X \mid Y, W, I=0) \mid W, I=0] \\ &= \operatorname{var}[E(X \mid Y, W, I=1) \mid W, I=0] + E[\operatorname{var}(X \mid Y, W, I=1) \mid W, I=0] \\ &= \operatorname{var}[Y\beta_Y + \mu\beta_W \mid W, I=0] + E[v^2 \mid W, I=0] = \sigma_0^2\beta_Y^2 + v^2\end{aligned}$$

$$\Leftrightarrow \sigma_0^2 = \frac{\varrho_0^2 - v^2}{\beta_Y^2}, \text{ assuming } \beta_Y \neq 0 .$$

❑

To be able to draw from the distribution $f(Y \mid W, I=0)$ normality is assumed.

Assumption B

$$f(Y \mid W, I=1) \sim N(\mu\Phi_1, \sigma_1^2) \tag{6.36}$$

$$f(Y \mid W, I=0) \sim N(\mu\Phi_0, \sigma_0^2) . \tag{6.37}$$

The validity of these assumptions is briefly assessed for LFS data. However, since only respondent data is available in the LFS sample only the assumptions made in (6.36) can be evaluated. In the LFS application we use $\ln(Y)$ and refer to the model $f(\ln(Y)\,|\,W, I=1)$. The table of coefficients, analysis of variance and diagnostic plots are given in appendix A6.4, applying the model to the LFS quarter March-May 2000. To evaluate the assumption of constant variance of the residuals the plot of the residuals versus fitted values, given in figure A6.4.1, is investigated. The plot shows a random spread with no obvious pattern indicating non-constant variance. The effect of the NMW level in the variable Y can be observed. The histogram of the residuals, given in figure A6.4.2, and the normal probability plot in figure A6.4.3 in the appendix show a distribution which approximates the normal distribution reasonably well. We therefore conclude that the assumptions made in (6.36), using the ln transformation on Y, seem adequate.

Under assumption B it follows for $\lambda(Y\,|\,W)$ that

$$\lambda(Y\,|\,W) = c_1 \exp\{c_2 Y^2 + c_3 Y\}, \tag{6.38}$$

where

$$c_1 = \frac{\sigma_1}{\sigma_0}\exp\{(2\sigma_0^2\sigma_1^2)^{-1}(\sigma_0^2(\mu\Phi_1)^2 - \sigma_1^2(\mu\Phi_0)^2\},$$

$$c_2 = (2\sigma_0^2\sigma_1^2)^{-1}(\sigma_0^2 - \sigma_1^2) \quad \text{and} \quad c_3 = (\sigma_0^2\sigma_1^2)^{-1}(\sigma_1^2\mu\Phi_0 - \sigma_0^2\mu\Phi_1).$$

Proof

$$\lambda(Y\,|\,W) = \frac{f(Y\,|\,W, I=0)}{f(Y\,|\,W, I=1)} = \frac{(2\pi\sigma_0^2)^{-1/2}\exp\{\dfrac{-1/2(Y-\mu\Phi_0)^2}{\sigma_0^2}\}}{(2\pi\sigma_1^2)^{-1/2}\exp\{\dfrac{-1/2(Y-\mu\Phi_1)^2}{\sigma_1^2}\}}$$

$$= \frac{\sigma_1}{\sigma_0} \exp\{(2\sigma_0^2\sigma_1^2)^{-1}[(\sigma_0^2 - \sigma_1^2)Y^2 + (2\sigma_1^2\mu\Phi_0 - 2\sigma_0^2\mu\Phi_1)Y + (\sigma_0^2(\mu\Phi_1)^2 - \sigma_1^2(\mu\Phi_0)^2]\}$$

$$= \frac{\sigma_1}{\sigma_0} \exp\{(2\sigma_0^2\sigma_1^2)^{-1}(\sigma_0^2(\mu\Phi_1)^2 - \sigma_1^2(\mu\Phi_0)^2)\}$$

$$* \exp\{(2\sigma_0^2\sigma_1^2)^{-1}(\sigma_0^2 - \sigma_1^2)Y^2 + (\sigma_0^2\sigma_1^2)^{-1}(\sigma_1^2\mu\Phi_0 - \sigma_0^2\mu\Phi_1)Y\}$$

$$= c_1 \exp\{c_2 Y^2 + c_3 Y\} \qquad \square$$

Under assumption B $\lambda(Y \mid W)$ can be expressed as a simple function of Y. An estimate $\hat{\lambda}(Y \mid W)$ can be obtained using estimates of the parameters $\sigma_0^2, \sigma_1^2, \Phi_0$ and Φ_1. This allows us to calculate $\lambda(\hat{y}_{j1} \mid \mathrm{w}_j,), ..., \lambda(\hat{y}_{jQ} \mid \mathrm{w}_j)$ and to find imputed values $\hat{y}_j$ for all $j \in \bar{r}$ using the weighted bootstrap method.

An alternative imputation method that can be defined under the above assumptions is the following. Under the assumptions A and B as well as the derivations in (6.34) and (6.35) it is possible to impute the missing values for Y by specifying an estimate of the distribution $f(Y \mid W, I = 0)$. Instead of drawing the imputed values from the estimated distribution $\hat{f}(Y \mid X, W, I = 0)$ they are drawn from the estimated distribution $\hat{f}(Y \mid W, I = 0)$. The mean $\mu\Phi_0$ and variance σ_0^2 can be estimated using estimates of observable parameters as given in (6.34) and (6.35). Assumption B says that $f(Y \mid W, I = 0) \sim N(\mu\Phi_0, \sigma_0^2)$. According to this assumption we can obtain imputed values for $j \in \bar{r}$ of the form

$$\hat{y}_j = \mu_j\hat{\Phi}_0 + \hat{\varepsilon}_j, \text{ where } \hat{\varepsilon}_j \sim N(0, \hat{\sigma}_0^2), \tag{6.39}$$

which leads to values $\hat{y}_j$ of the form

$$\hat{y}_j = \frac{\hat{\delta}_{00} - \hat{\beta}_{0W}}{\hat{\beta}_Y} + \frac{\hat{\delta}_{10} - \hat{\beta}_{1W}}{\hat{\beta}_Y}\mathrm{w}_{1j} + ... + \frac{\hat{\delta}_{V0} - \hat{\beta}_{VW}}{\hat{\beta}_Y}\mathrm{w}_{Vj} + \hat{\varepsilon}_j, \tag{6.40}$$

where $\hat{\beta}_W = (\hat{\beta}_{0W}, ..., \hat{\beta}_{VW})'$ and $\hat{\delta}_{00} = (\hat{\delta}_{00}, ..., \hat{\delta}_{V0})'$. This imputation method will be referred to as *derived imputation*. A disadvantage of this derived imputation method is its strong dependency on correct model assumptions, in particular the correct specification of the parameters Φ_0 and σ_0^2. It is expected that this method is sensitive to model misspecification. The derived imputation

method is not a hot deck method, whereas the weighted bootstrap imputation method can be defined as a hot deck method, e.g. using hot deck imputation within classes or nearest neighbour imputation to sample from the distribution $f(Y \mid X, W, I=1)$. It is hoped that this makes the weighted bootstrap imputation method slightly less sensitive to model misspecification, since the assumptions about $f(Y \mid X, W, I=1)$ are relaxed and the method depends on the estimation of $\lambda(Y \mid W)$ via the rejection algorithm. An indication that the derived method is more sensitive to model misspecifications can be seen in the results of the simulation study, e.g. in table 6.13.

6.3.4.1 Theoretical Investigation of the Weighted Bootstrap Imputation Method

In the following some of the properties of the weighted bootstrap imputation method are investigated theoretically. It is assumed that the parameters Φ_0 and σ_0^2 are correctly specified, using the derivations in (6.34) and (6.35), and that assumption B holds.

a.) assuming $\sigma_0^2 = \sigma_1^2$:

To illustrate the properties of the additional term $\lambda(Y \mid W)$ we first consider the simpler case where $\sigma_0^2 = \sigma_1^2$. The term $\lambda(Y \mid W)$, used to generate the probability of selection for $\hat{y}_{jq}^*$, allocates simply speaking a greater probability to values $\hat{y}_{jq}^*$ that are 'close' to $\mu\Phi_0$ and smaller probabilities to values 'close' to $\mu\Phi_1$ (assuming $\mu\Phi_0 \neq \mu\Phi_1$). For the case where $\sigma_0^2 = \sigma_1^2$, it follows that

$$\lambda(Y \mid W) = \frac{\exp\{\frac{-1/2(Y-\mu\Phi_0)^2}{\sigma_1^2}\}}{\exp\{\frac{-1/2(Y-\mu\Phi_1)^2}{\sigma_1^2}\}} \tag{6.41}$$

$$= \exp\{(2\sigma_1^2)^{-1}[(Y-\mu\Phi_1)^2 - (Y-\mu\Phi_0)^2]\} \tag{6.42}$$

$$= \exp\{\frac{Y(2\mu\Phi_0 - 2\mu\Phi_1) + c}{2\sigma_1^2}\}, \text{ where } c = \mu\Phi_1^2 - \mu\Phi_0^2.$$

Note that per definition $Y \geq 0$. (In fact we found in the simulation study and in the application to LFS data, discussed in sections 6.3.4.3 and 6.3.4.4 respectively, that $\ln(\hat{y}_{jq}^*) \geq 0$ for all $j \in \bar{r}$ and

$q = 1,..., Q$.) We assume that $\mu\Phi_1, \mu\Phi_0 \geq 0$. Interpreting equation (6.42) we can see that in general, if $\hat{y}^*_{jq}$ is 'close' to $\mu\Phi_1$, $\lambda(\hat{y}^*_{jq} \mid W)$ is small, whereas if $\hat{y}^*_{jq}$ is close to $\mu\Phi_0$, $\lambda(\hat{y}^*_{jq} \mid W)$ is large and the associated probability $\hat{p}^*_{jq}$ for selecting this value $\hat{y}^*_{jq}$ is therefore large. In the case where $\mu\Phi_1 < \mu\Phi_0$, $\lambda(Y \mid W)$ is of the form as indicated in figure 6.6, and $\lambda(Y \mid W)$ is small if Y is close to zero or close to $\mu\Phi_1$ and large if Y is close to $\mu\Phi_0$. However, a large value of Y, even if it is not close to $\mu\Phi_0$ has also a large value $\lambda(Y \mid W)$ and is therefore likely to be selected for imputation. In this case we have $(Y - \mu\Phi_1)^2 \gg (Y - \mu\Phi_0)^2$ and therefore in (6.41)

$$\exp\{\frac{-1/2(Y-\mu\Phi_1)^2}{\sigma_1^2}\} \ll \exp\{\frac{-1/2(Y-\mu\Phi_0)^2}{\sigma_1^2}\},$$

i.e. the denominator is much smaller than the numerator, which leads to a large value of $\lambda(Y \mid W)$. This can potentially lead to biased results when using the weighted bootstrap imputation method and needs to be accounted for when applying this method.

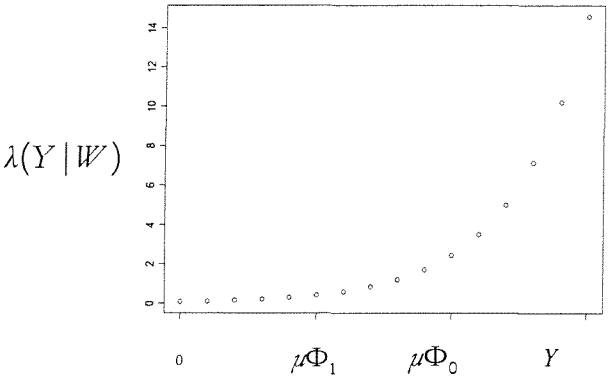

Figure 6.6:
Form of $\lambda(Y \mid W)$,
if $\sigma_0^2 = \sigma_1^2$ and $\mu\Phi_1 < \mu\Phi_0$.

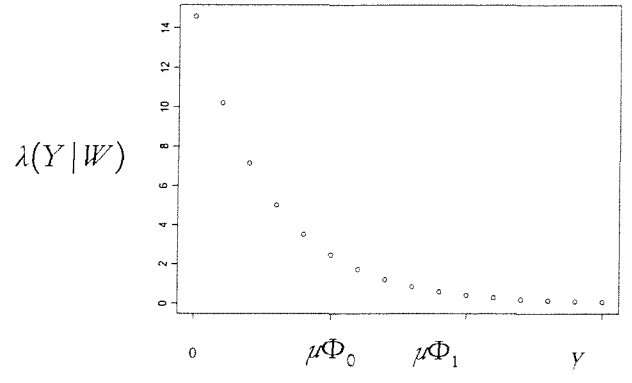

Figure 6.7:
Form of $\lambda(Y \mid W)$,
if $\sigma_0^2 = \sigma_1^2$ and $\mu\Phi_0 < \mu\Phi_1$.

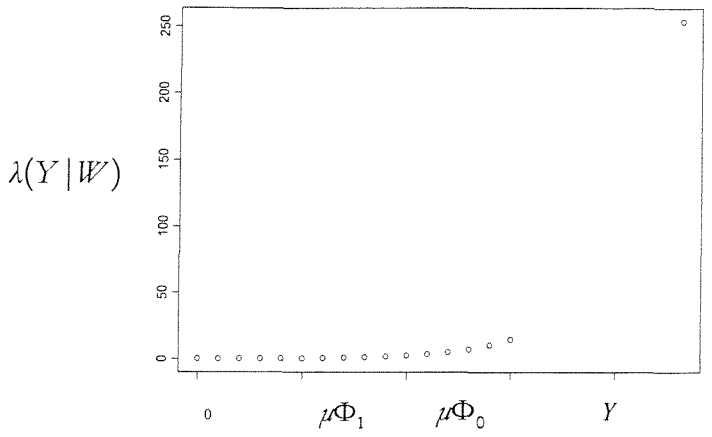

Figure 6.8: Form of $\lambda(Y \mid W)$, if $\sigma_0^2 = \sigma_1^2$ and $\mu\Phi_1 < \mu\Phi_0$, with outlying value.

In the case where $\mu\Phi_0 < \mu\Phi_1$, $\lambda(Y \mid W)$ is of the form as shown in figure 6.7 and large values for $\lambda(Y \mid W)$ are obtained if Y is close to zero. Again, it seems sensible to adjust for large values of $\lambda(Y \mid W)$ to avoid imputing small values of Y disproportionately often. A potential disadvantage of the weighted bootstrap imputation method is that the computed probability of selection for imputation, $\hat{p}_{j1}^*$, strongly depends on the specific values $\hat{y}_{j1}^*, \ldots, \hat{y}_{jQ}^*$, selected as donor values. Whereas in the case of rejection sampling, the probability $\hat{p}_{jq}^*$ only depends on $\hat{y}_{jq}^*$. Therefore, if already one value $\hat{y}_{jq^*}^*$ is 'very different' from the other selected values for imputation this value might get either a very small or a very large probability of selection. In this case the form of $\lambda(Y \mid W)$ could take on a form as indicated in figure 6.8, where only one outlying value influences the form of $\lambda(Y \mid W)$ immensely. The occurrence of a very large probability should be avoided using appropriate adjustment methods as discussed later.

Under correct model specifications the problem described here might not have a great impact since if we account for the covariates W only possible values $\hat{y}_{jq}^*$ should be obtained for imputation that do not have this property of differing 'very much' from the other selected values for imputation. However, under misspecification of model assumptions we might encounter the problem of selecting such an outlying value $\hat{y}_{jq^*}^*$, where $\lambda(\hat{y}_{jq^*}^* \mid W) \gg \lambda(\hat{y}_{jq}^* \mid W)$ for $q \in \{1, \ldots, Q\}$. This could be the case for $\hat{y}_{jq^*}^*$ if the predicted value and the observed value from

$f(Y \mid X, W, I=1)$ differ much, i.e. in the case of large residuals, since under predictive mean matching imputation the donor values are selected based on the predicted values.

b.) assuming $\sigma_0^2 \neq \sigma_1^2$:

In the case where $\sigma_0^2 \neq \sigma_1^2$, $\lambda(Y \mid W)$ is of the form as specified in (6.38). This equation has a

minimum at $Y = -c_3/(2c_2)$ if $\sigma_0 > \sigma_1$ and a

maximum at $Y = -c_3/(2c_2)$ if $\sigma_0 < \sigma_1$.

Proof

$\lambda'(Y \mid W) = c_1 \exp\{c_2 Y^2 + c_3 Y\} * (2c_2 Y + c_3) = 0$ for $Y = -c_3/(2c_2)$, assuming $\sigma_0^2 \neq \sigma_1^2$.

$\lambda''(Y \mid W) = c_1 \exp\{c_2 Y^2 + c_3 Y\} * 2c_2$,

which is greater than zero if $c_1 c_2 > 0$ and smaller than zero if $c_1 c_2 < 0$. Assuming $\sigma_1, \sigma_0 \geq 0$ we have $c_1 > 0$ always and $c_2 < 0$ if $\sigma_0 < \sigma_1$ and $c_2 > 0$ if $\sigma_0 > \sigma_1$. Therefore $\lambda(Y \mid W)$ has a minimum if $\sigma_0 > \sigma_1$ and a maximum if $\sigma_0 < \sigma_1$.

❑

The extreme value is therefore of the form

$$\frac{-c_3}{2c_2} = \frac{-(\sigma_0^2\sigma_1^2)^{-1}(\sigma_1^2\mu\Phi_0 - \sigma_0^2\mu\Phi_1)}{(\sigma_0^2\sigma_1^2)^{-1}(\sigma_0^2 - \sigma_1^2)} = \frac{\sigma_0^2\mu\Phi_1 - \sigma_1^2\mu\Phi_0}{(\sigma_0^2 - \sigma_1^2)}. \qquad (6.43)$$

In the case where $\sigma_0 > \sigma_1$, $\lambda(Y \mid W)$ is of the form as indicated in figure 6.9. A value $\hat{y}^*_{jq^*}$, for which $d^*_{jq^*} = \max_q |\, y^*_{jq^*} - (-c_3/(2c_2))|$, $q = \{1, ..., Q\}$, obtains the largest value for $\lambda(Y \mid W)$ and therefore the largest probability of selection. Again, an adjustment against large values of $\lambda(Y \mid W)$ seems necessary if weighted bootstrap imputation is applied in practice. Taking into account X and W, the draws from $f(Y \mid X, W, I=1)$ obtaining $\hat{y}^*_{j1}, ..., \hat{y}^*_{jQ}$ should all be within a certain range, such that unusual large values of $\lambda(Y \mid W)$ should be avoided. However, especially under model misspecification we might select one (or more) $\hat{y}^*_{jq^*}$ for which the resulting value for

$\lambda(Y \mid W)$ is large. The probability of selection for imputation $\hat{p}^*_{jq^*}$ is then close to one, such that it is likely that this outlying value $\hat{y}^*_{jq^*}$ is used for imputation. This issue needs to be addressed when applying the weighted bootstrap imputation method in practice.

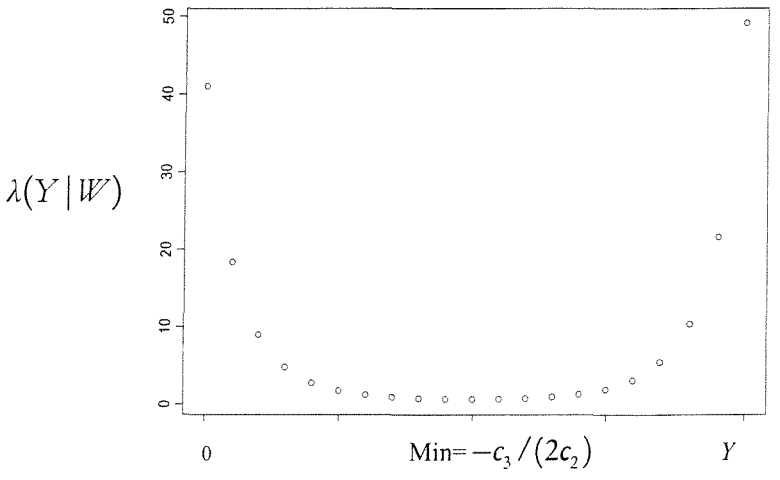

Figure 6.9: Form of $\lambda(Y \mid W)$, if $\sigma_0 > \sigma_1$.

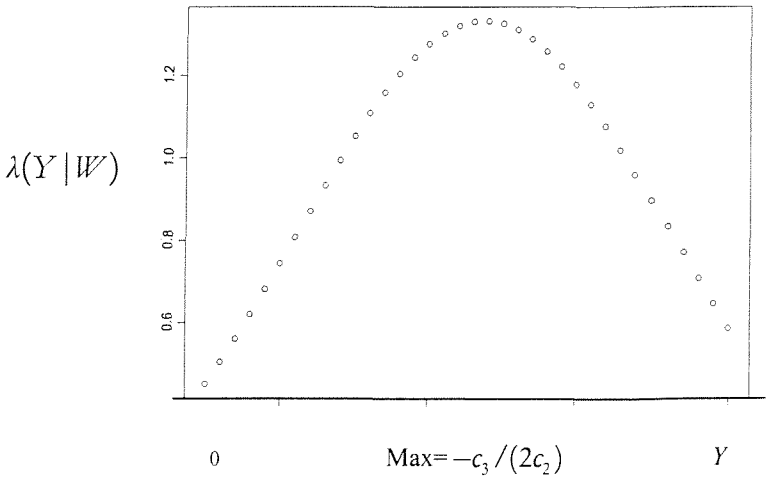

Figure 6.10: Form of $\lambda(Y \mid W)$, if $\sigma_0 < \sigma_1$.

Figure 6.10 indicates the form of $\lambda(Y|W)$ in the case where $\sigma_0 < \sigma_1$. Here, the values for $\lambda(Y|W)$ for outlying values of Y, i.e. large or small values of Y, are small. In this special case an adjustment might not be necessary. The form of $\lambda(\ln(Y)|W)$ therefore depends on the specifications of the parameters, such as σ_0^2 and σ_1^2, which might be a robustness concern. Note that applying the weighted bootstrap imputation method to LFS data using $\ln(Y)$ and $\ln(X)$ in the regression models we found that $\hat{\sigma}_0 > \hat{\sigma}_1$, where $\hat{\sigma}_0^2 = 0.145$ and $\hat{\sigma}_1^2 = 0.065$. In the following we will therefore mostly concentrate on the case where $\sigma_0 > \sigma_1$.

6.3.4.2 Possible Adjustment Methods

In the above analysis we have seen that the generation of disproportionately large probabilities and therefore the generation of disproportionately large values of $\lambda(Y|W)$ should be avoided. This might be necessary even if the underlying assumptions are correc. Three possible adjustment methods are presented.

a.) Adjustment against a small denominator in $\lambda(Y|W)$, setting

$$f(Y|W,I=1) = \max\{f(Y|W,I=1),\omega\},$$

where ω is a constant, $\omega \in \mathbb{R}^+$. This constant is chosen to be ω = 1/10, 1/100, 1/1000.

b.) Adjustment against large values of $\lambda(Y|W)$, restricting the value for $\lambda(Y|W)$, i.e.

$$\lambda(\hat{y}^*_{jq}|W) = \min\{\lambda(\hat{y}^*_{jq}|W),\omega\} \quad \forall\, q = 1,\ldots, Q \text{ and } j \in \bar{r},$$

where for example ω =1.5, 5,10.

c.) Adjustment against large values of $\lambda(Y|W)$, where values $\hat{y}^*_{jq^*}$ for which $\lambda(\hat{y}^*_{jq^*}|W)$ are large are excluded from the set of possible donor values, $\hat{y}^*_{j1}, \ldots, \hat{y}^*_{jQ}$. This could be done by excluding $\hat{y}^*_{jq^*}$ for which $d^*_{jq^*} = \max_q |\hat{y}^*_{jq} - (\frac{-c_3}{2c_2})|$. However, such a method might introduce additional bias.

So far we have assumed that the assumptions made in the derivation of the weighted bootstrap imputation method hold. These are that the parameters Φ_0 and σ_0^2 are correctly specified and that the normal assumptions in B hold. It is also of interest to investigate the properties of the

weighted bootstrap imputation method under model misspecification. This is addressed empirically in the following simulation study. A general problem is the correct specification of estimates of the parameters Φ_0 and σ_0^2. For example $\hat{\sigma}_0^2$ could be negative because of the way it is derived. However, in the application to the LFS as well as in the simulation studies considered here this case did not occur. Further investigation has shown, however, that the estimators of Φ_0 and σ_0^2 might be biased because of the way they are derived using (6.34) and (6.35). The assumption of normality made in assumption B was investigated, as discussed earlier, based on respondent data and using $\ln(Y)$ as the dependent variable. This analysis showed that the assumption of normality may hold approximately. Therefore, a disadvantage of the weighted bootstrap imputation method are possible misspecifications of the parameters Φ_0 and σ_0^2 when applied in practice. Applying the weighted bootstrap imputation method in a simulation study under model misspecification or to LFS data an increase in the occurrence of unusually large values of $\lambda(\ln(Y)|W)$ can be observed. Making adjustments against large values of $\lambda(\ln(Y)|W)$ will therefore be necessary to make the weighted bootstrap imputation method more robust to model misspecification.

6.3.4.3 Simulation Study

To evaluate the performance of the weighted bootstrap imputation method and to compare this approach to the MAR based imputation methods the results of a simulation study are presented. The simulation study is carried out based on the assumptions made above, i.e. the assumption of CME and assumptions A and B. The simulation study is carried out (a) not taking into account any covariates W and (b) including covariates W, which involves bootstrapping one LFS sample and generating values for $\ln(Y)$, $\ln(X)$, and I for all $i \in s$. In the following simulation study the weighted bootstrap imputation method is based on using $\ln(Y)$ and $\ln(X)$.

a.) Absence of Covariates W in the Simulation Study

First we analyse the special case where no covariates are present. The variables we need to obtain are therefore $\ln(Y)$, $\ln(X)$ and I, for I=1 and I=0. To generate the data in sample $s^{(a)}$ the probability of response on the variable Y, $pr(I_i = 1)$, is fixed to φ for all i. This is estimated from LFS data to be $\hat{\varphi} = 0.43$. The number of units in each sample is fixed to be $n = 10{,}000$. We

have $E_R(n_r) = n\varphi$ and $E_R(n_{\bar{r}}) = n(1-\varphi)$. This generates the indicator of response I. According to the assumptions A and B values for $\ln(Y)$ are drawn from the normal distributions $\ln(y_i) \sim N(\hat{\Phi}_1, \hat{\sigma}_1^2)$ for $I_i = 1$ and $\ln(y_i) \sim N(\hat{\Phi}_0, \hat{\sigma}_0^2)$ for $I_i = 0$. The parameters Φ_1 and σ_1^2 are estimated from LFS data. Estimates of the unknown parameters Φ_0 and σ_0^2 are derived using (6.34) and (6.35), where the parameters in these derivations are also estimated from LFS data. The variable $\ln(X)$ is generated conditional on the variables $\ln(Y)$ and I, using the assumption of CME. Since $E(\ln(X) \mid \ln(Y), I = 0) = E(\ln(X) \mid \ln(Y), I = 1) = \beta_0 + \beta_Y \ln(Y)$, where β_0 and β_Y are estimated from LFS data, we obtain predicted values for $\ln(X)$. Adding on random residuals $\hat{\varepsilon}_i$, where $\hat{\varepsilon}_i \sim N(0, \hat{\nu}^2)$ and $\hat{\nu}^2 = \hat{\text{var}}(\ln(X) \mid \ln(Y), I = 0) = \hat{\text{var}}(\ln(X) \mid \ln(Y), I = 1)$ estimated from LFS data, gives the variable $\ln(X)$ in $s^{(a)}$. The nonresponse mechanism follows the assumption of CME nonresponse.

The imputation is carried out according to the weighted bootstrap imputation method as described in (6.29). For $j \in \bar{r}$ a value $\ln(\hat{y}_j^*)$, is drawn, where $\ln(\hat{y}_j^*) \sim \hat{f}(\ln(Y) \mid \ln(X), I = 1)$. This is done by obtaining the predicted value, $\hat{\alpha}_{01} + \hat{\alpha}_{X1} \ln(x_j)$, and adding on a random residual $\hat{\varepsilon}_j \sim N(0, \hat{\pi}_1^2)$, where the estimates $\hat{\alpha}_{01}$, $\hat{\alpha}_{X1}$ and $\hat{\pi}_1^2$ are obtained using the observed data in sample $s^{(a)}$ ($\hat{\alpha}_{01}$ denotes the intercept of the observed distribution $\hat{f}(\ln(Y) \mid \ln(X), I = 1)$ and $\hat{\alpha}_{X1}$ the coefficient of $\ln(X)$). The value of $\hat{\lambda}(\ln(Y))$ is obtained using the expression (6.38), $\lambda(\ln(\hat{y}_j^*)) = c_1 \exp\{c_2 \ln(\hat{y}_j^*)^2 + c_3 \ln(\hat{y}_j^*)\}$, where all parameters contained in c_1, c_2 and c_3 are estimated using sample $s^{(a)}$. Q values are drawn for person j from $\hat{f}(\ln(Y) \mid \ln(X), W, I = 1)$, denoted $\ln(\hat{y}_{j1}^*), \ldots, \ln(\hat{y}_{jQ}^*)$, and $\hat{\lambda}(\ln(\hat{y}_{j1}^*)), \ldots, \hat{\lambda}(\ln(\hat{y}_{jQ}^*))$ are obtained. The finally imputed value for $j \in \bar{r}$, denoted $\ln(\hat{y}_j^*)$, is sampled out of the Q possible values $\ln(\hat{y}_{j1}^*), \ldots, \ln(\hat{y}_{jQ}^*)$ with probabilities

$$\hat{p}_{jq}^* = \hat{\lambda}(\ln(\hat{y}_{jq}^*)) / \sum_{q=1}^{Q} \hat{\lambda}(\ln(\hat{y}_{jq}^*)) \text{ for } q = \{1, \ldots, Q\}. \tag{6.44}$$

Different values for Q are explored. In total, A =1000 iterations are used for the simulation.

The imputation method, imputing a value obtained from the predicted value for the variable $\ln(Y)$, $\hat{\alpha}_{01} + \hat{\alpha}_{X1} \ln(x_j)$, adding on Q random residuals drawn from a normal distribution and accepting one out of these Q values, $\ln(\hat{y}_{j1}^*), \ldots, \ln(\hat{y}_{jQ}^*)$, with probability $\hat{p}_{jq}^*$ will be referred to as *weighted bootstrap random regression imputation*, abbreviated wboot-reg. imp.(1). The number in

parentheses indicates the total number of imputations for each nonrespondent $j \in \bar{r}$, which is 1 for most applications considered here. Alternatively, it is possible to draw $\ln(\hat{y}_j^*)$ from the estimated distribution $\hat{f}(\ln(Y) \mid \ln(X), I=1)$ using nearest neighbour imputation selecting the $Q/2$ nearest neighbours above and below the predicted value for the nonrespondent (in the same way as described in chapter 5) or using hot deck imputation within classes selecting Q donors for each nonrespondent with replacement out of the appropriate imputation class (in the same way as described in chapter 3) and selecting one donor out of the possible Q donors according to $\hat{p}_{jq}^*$. These methods will be referred to as *weighted bootstrap nearest neighbour imputation* (wboot-NN(1)) and *weighted bootstrap hot deck imputation* (wboot-HDIwr(1)), respectively. The value for Q specifies the size of the set of possible donors for each nonrespondent. The point estimators $\hat{P}_1.$ and $\hat{P}_2.$ are obtained as defined in (6.32), where the probabilities $\hat{p}_{jq}$ are as defined in (6.44). For comparison, the *derived imputation method* is also applied drawing possible values for the missing $\ln(Y)$'s from $f(\ln(Y) \mid W, I=0)$ based on the assumption A and B as well as the derivations in (6.34) and (6.35).

In addition, the results under the weighted bootstrap imputation method are compared to random regression imputation (reg.imp(1)), hot deck imputation within classes (HDIwr(1)) and nearest neighbour imputation (NN(1)) without taking into account $\lambda(\ln(Y))$. Here only single value imputation is used, i.e. for all $j \in \bar{r}$ one value is imputed. These methods are based on the estimated distribution $\hat{f}(\ln(Y) \mid \ln(X), I=1)$ as it would be appropriate under the assumption of MAR. The results are given in table 6.6. Overall, we can conclude that under the conditions specified here the weighted bootstrap imputation method performs well with most relative biases under 2%. We can see that the bias tends to zero if $Q \to \infty$. There is some indication that the larger Q is the smaller is the variance of the estimator. The smallest simulation variance is found for wboot-reg.imp(1). The method wboot-HDIwr(1) seems to perform slightly worse with respect to bias than regression and nearest neighbour imputation, which might be related to the use of imputation classes. We also see a good performance of the derived imputation method with relative biases close to zero, as it is expected since the necessary conditions are fulfilled and estimates of the parameters Φ_0 and σ_0^2 are likely to be well specified. The methods, valid under the assumption of MAR, are, as expected, significantly biased, with a bias of around 5%-6% for $\hat{P}_1.$ and around 12%-14% for $\hat{P}_2.$.

Imputation Method	Bias of $\hat{P}_1$.	Rel. Bias $\hat{P}_1$.	Bias of $\hat{P}_2$.	Rel. Bias $\hat{P}_2$.	$V(\hat{P}_1.)$	$MSE(\hat{P}_1.)$
Wboot-reg. imp.(1) Q=6	$9.88*10^{-4}$ $(2.95*10^{-4})^*$	1.41%	$3.74*10^{-3}$ $(3.85*10^{-4})^*$	2.35%	$8.75*10^{-6}$	$8.84*10^{-6}$
Wboot-reg. imp.(1) Q=10	$6.32*10^{-4}$ $(2.91*10^{-4})^*$	0.90%	$2.20*10^{-3}$ $(3.90*10^{-4})^*$	1.38%	$8.48*10^{-6}$	$8.52*10^{-6}$
Wboot- reg. imp.(1) Q=20	$4.09*10^{-4}$ $(2.90*10^{-4})$	0.58%	$1.05*10^{-3}$ $(3.87*10^{-4})^*$	0.66%	$8.46*10^{-6}$	$8.47*10^{-6}$
Wboot- reg. imp.(1) Q=100	$0.33*10^{-4}$ $(2.74*10^{-4})$	0.04%	$0.34*10^{-3}$ $(3.79*10^{-4})^*$	0.02%	$7.52*10^{-6}$	$7.52*10^{-6}$
Wboot- NN(1) Q=6	$8.06*10^{-4}$ $(3.21*10^{-4})^*$	1.15%	$3.95*10^{-3}$ $(4.92*10^{-4})^*$	2.48%	$10.3*10^{-6}$	$10.3*10^{-6}$
Wboot- NN(1) Q=10	$4.30*10^{-4}$ $(3.20*10^{-4})$	0.61%	$2.29*10^{-3}$ $(4.81*10^{-4})^*$	1.44%	$10.2*10^{-6}$	$10.3*10^{-6}$
Wboot- NN(1) Q=20	$-1.60*10^{-4}$ $(3.14*10^{-4})$	-0.22%	$1.11*10^{-3}$ $(4.85*10^{-4})^*$	0.69%	$10.2*10^{-6}$	$9.86*10^{-6}$
Wboot- HDIwr(1) Q=6	$8.54*10^{-4}$ $(3.16*10^{-4})^*$	1.22%	$4.60*10^{-3}$ $(4.96*10^{-4})^*$	2.89%	$9.98*10^{-6}$	$10.0*10^{-6}$
Wboot- HDIwr(1) Q=10	$8.38*10^{-4}$ $(3.13*10^{-4})^*$	1.19%	$2.99*10^{-3}$ $(4.80*10^{-4})^*$	1.88%	$9.84*10^{-6}$	$9.91*10^{-6}$
Wboot- HDIwr(1) Q=20	$5.07*10^{-4}$ $(3.19*10^{-4})$	0.72%	$1.03*10^{-3}$ $(4.86*10^{-4})^*$	0.64%	$9.84*10^{-6}$	$10.2*10^{-6}$
Derived imputation(1)	$2.13*10^{-4}$ $(3.06*10^{-4})$	0.30%	$-3.40*10^{-5}$ $(3.92*10^{-4})$	-0.02%	$9.42*10^{-6}$	$9.42*10^{-6}$
Reg. imp.(1)	$4.15*10^{-3}$ $(3.09*10^{-4})^*$	5.93%	$2.25*10^{-2}$ $(4.33*10^{-4})^*$	14.13%	$9.55*10^{-6}$	$11.2*10^{-6}$
NN(1)	$4.12*10^{-3}$ $(3.39*10^{-4})^*$	5.89%	$2.29*10^{-2}$ $(5.56*10^{-4})^*$	14.42%	$11.5*10^{-6}$	$13.2*10^{-6}$
HDIwr(1)	$-2.88*10^{-3}$ $(2.15*10^{-4})^*$	-4.14%	$1.97*10^{-2}$ $(2.90*10^{-4})^*$	12.37%	$4.62*10^{-6}$	$5.45*10^{-6}$

Table 6.6: Results for the weighted bootstrap imputation method for different values of Q, based on random regression, nearest neighbour and hot deck imputation, under the assumption of CME. For comparison the results of the derived imputation and MAR-based random regression, nearest neighbour and hot deck imputation are shown.

b.) Including Covariates W in the Simulation Study

In the following simulation study covariates W are included. The sampling procedure is simulated by bootstrapping an original data set with replacement from the LFS assuming an infinite population. In total A =1000 bootstrap samples, denoted $s^{(a)}$, $a = 1,...,A$, of size n=10,000 are generated. The information for the variables W are obtained by this bootstrapping process. The variable $\ln(Y)$ is generated for each $s^{(a)}$ using the predictions of a linear model, which takes into account the derived variable $\ln(X)$ and other covariates contained in $s^{(a)}$ and adding on random residuals. To avoid replication of units the variable $\ln(X)$ is also generated using a linear regression model with added residuals as described in section 3.3. All necessary parameters are specified using estimates from the LFS. Nonresponse is generated according to the assumption of CME, such that $X \perp I | (Y,W)$ using the same approach described in section 6.2.2. Note that the normal assumptions as made in assumption B are approximately valid because of the way the data $\ln(Y)$ is generated. The weighted bootstrap method is applied using the formula (6.38) to obtain $\lambda(\ln(Y)|W)$. Estimates of the parameters in c_1, c_2 and c_3 are obtained in each iteration a based on data from sample $s^{(a)}$. As before in case a.) we draw Q values for person j from $f(\ln(Y)|\ln(X),W,I=1)$ and obtain $\lambda(\ln(\hat{y}^*_{j1})),...,\lambda(\ln(\hat{y}^*_{jQ}))$. The finally imputed value for $j \in \bar{r}$, $\ln(\hat{y}^*_j)$, is sampled out of the Q possible values $\ln(\hat{y}^*_{j1}),...,\ln(\hat{y}^*_{jQ})$ with probabilities $\hat{p}^*_{jq}$ for $q=\{1,...,Q\}$. The weighted bootstrap method is carried out based on drawing $\ln(\hat{y}^*_j) \sim \hat{f}(\ln(Y)|\ln(X),W,I=1)$ using either random regression imputation, hot deck imputation within classes or nearest neighbour imputation. We refer to these methods as wboot-reg. imp.(1), wboot-HDIwr(1) and wboot-NN(1). For comparison, the results under the derived imputation method as well as MAR-based random regression, nearest neighbour and hot deck imputation within classes are presented.

Results of the Simulation Study Including Covariates W

b.1) Performance of the Weighted Bootstrap Imputation Method under Ideal Conditions

Here, the same covariates are used consistently throughout the simulation study, i.e. the generation of $\ln(X)$ and $\ln(Y)$, the model to generate nonresponse according to the assumption of CME as

well as all imputation models are based on the same covariates, denoted W.[1] Later, the case is discussed where the covariates used for the generation of $\ln(Y)$ and the generation of $\ln(X)$ as well as the model for nonresponse differ from the covariates W that are used within the imputation process to test the robustness of the weighted bootstrap imputation method under model misspecification. The point estimators of interest $\hat{P}_1.$ and $\hat{P}_2.$ are obtained as defined in (6.32) and (6.33), where the probabilities $\hat{p}_{jq}$ are as defined in (6.44), based on $\ln(Y)$.

Analysing the results in table 6.7 we can see that even under ideal conditions the weighted bootstrap method shows a significant bias of around 2% for $\hat{P}_1.$ for regression and nearest neighbour imputation. The relative bias for $\hat{P}_2.$ is under 1%. All biases are significant. The derived imputation method shows similar results. Weighted bootstrap imputation based on HDI within classes shows for both point estimators a higher bias. For $\hat{P}_1.$ the relative bias is around 3.5% and for $\hat{P}_2.$ it is around 1.5%. The larger bias for the wboot-HDIwr method seems to be caused by the occurrence of some larger values of $\lambda(\ln(Y)|W)$. An investigation of the values for $\lambda(\ln(Y)|W)$ under this method shows that large values of $\lambda(\ln(Y)|W)$ seem to occur more frequently for HDI in comparison to the other methods. This might be caused by the imputation within classes where the donor values within each class cover a wider range of possible donor values as for example in comparison to the nearest neighbour method, where possible donor values are chosen to be close to the predicted value of the nonrespondent. The nearest neighbour method effectively selects donors from a tightly defined class around each nonrespondent. This means that the form of the weighted bootstrap method using HDIwr might be less attractive than wboot-NN imputation. Note that with increasing Q the bias can be reduced, however it is still significant even if Q is large. Table 6.7 also shows the performance of MAR-based random regression, HDI within classes and nearest neighbour imputation, using just one imputation. The relative bias is between 5.6% to 7.4%. That means that taking into account the value of $\lambda(\ln(Y)|W)$ instead of just sampling from $f(\ln(Y)|\ln(X),W,I=1)$ reduces, as expected, the bias under the assumption of CME.

[1] These covariates W are: major occupation group (SOC), part-time (PT), qualification (Q), Age and Age squared, length of time continuously employed (EMPMON), head of household (HOH), married, number of employees at work place (SIZE), industry section (IND), region (REG) and gender (FEMALE).

Imputation Method	Bias of $\hat{P}_1.$	Rel. Bias $\hat{P}_1.$	Bias of $\hat{P}_2.$	Rel. Bias $\hat{P}_2.$	$V(\hat{P}_1.)$	$MSE(\hat{P}_1.)$
Wboot- reg. imp.(1) Q=6	$10.37*10^{-4}$ $(0.93*10^{-4})^*$	2.12%	$17.70*10^{-4}$ $(1.49*10^{-4})^*$	0.98%	$8.72*10^{-6}$	$9.79*10^{-6}$
Wboot-reg. imp.(1) Q=10	$9.99*10^{-4}$ $(0.96*10^{-4})^*$	2.03%	$13.70*10^{-4}$ $(1.51*10^{-4})^*$	0.75%	$9.30*10^{-6}$	$10.30*10^{-6}$
Wboot-reg. imp.(1) Q=20	$10.31*10^{-4}$ $(0.96*10^{-4})^*$	2.09%	$10.13*10^{-4}$ $(1.54*10^{-4})^*$	0.56%	$9.24*10^{-6}$	$10.30*10^{-6}$
Wboot- NN(1) Q=6	$10.15*10^{-4}$ $(1.00*10^{-4})^*$	2.07%	$16.52*10^{-4}$ $(1.74*10^{-4})^*$	0.91%	$10.09*10^{-6}$	$11.13*10^{-6}$
Wboot-NN(1) Q=10	$9.99*10^{-4}$ $(1.01*10^{-4})^*$	2.03%	$12.88*10^{-4}$ $(1.75*10^{-4})^*$	0.71%	$10.37*10^{-6}$	$11.37*10^{-6}$
Wboot-NN(1) Q=20	$10.01*10^{-4}$ $(1.03*10^{-4})^*$	2.03%	$9.78*10^{-4}$ $(1.76*10^{-4})^*$	0.54%	$10.71*10^{-6}$	$11.72*10^{-6}$
Wboot- HDIwr(1) Q=6	$17.73*10^{-4}$ $(1.03*10^{-4})^*$	3.61%	$31.21*10^{-4}$ $(1.79*10^{-4})^*$	1.72%	$10.77*10^{-6}$	$13.91*10^{-6}$
Wboot-HDIwr(1) Q=10	$17.39*10^{-4}$ $(1.71*10^{-4})^*$	3.54%	$28.15*10^{-4}$ $(1.83*10^{-4})^*$	1.56%	$11.48*10^{-6}$	$14.50*10^{-6}$
Wboot-HDIwr(1) Q=20	$17.13*10^{-4}$ $(1.10*10^{-4})^*$	3.49%	$23.78*10^{-4}$ $(1.82*10^{-4})^*$	1.31%	$12.10*10^{-6}$	$15.03*10^{-6}$
Derived imputation(1)	$10.65*10^{-4}$ $(1.00*10^{-4})^*$	2.17%	$6.93*10^{-4}$ $(1.60*10^{-4})^*$	0.38%	$10.04*10^{-6}$	$11.18*10^{-6}$
Reg. imp.(1)	$30.56*10^{-4}$ $(0.79*10^{-4})^*$	6.55%	$98.93*10^{-4}$ $(1.59*10^{-4})^*$	5.48%	$6.26*10^{-6}$	$15.59*10^{-6}$
NN(1)	$26.36*10^{-4}$ $(0.87*10^{-4})^*$	5.65%	$105.5*10^{-4}$ $(1.63*10^{-4})^*$	5.85%	$7.61*10^{-6}$	$14.56*10^{-6}$
HDIwr(1)	$34.70*10^{-4}$ $(0.88*10^{-4})^*$	7.43%	$127.3*10^{-4}$ $(1.66*10^{-4})^*$	7.05%	$7.83*10^{-6}$	$19.88*10^{-6}$

Table 6.7: Results for the weighted bootstrap imputation method with Q=6, 10 and 20, based on regression imputation, nearest neighbour and hot deck imputation, under the assumption of CME nonresponse (including W). For comparison we use the derived imputation method as well as MAR-based random regression, nearest neighbour and hot deck imputation.

The weighted bootstrap fractional imputation method as specified in (6.33) is also investigated. As expected, the results for the biases are very similar to those using the weighted bootstrap method under single imputation. However, a gain in efficiency between 8%-12% is observed. (The results are not reported here.) In addition, possible adjustment methods are investigated to improve the results of the weighted bootstrap imputation method. As discussed above when using the weighted bootstrap method it is advised to restrict the denominator of $\lambda(\ln(Y)|W)$ or to restrict the size of $\lambda(\ln(Y)|W)$. First, the denominator is restricted, such that $f(\ln(Y)|W,I=1) = \max\{f(\ln(Y)|W,I=1),\omega\}$, where ω is chosen to be ω = 1/10, 1/100 and 1/1000. It is found that mild restrictions such as ω =1/100 or 1/1000 do not have a great impact on the reduction of the bias. For ω =1/10 a greater reduction in the bias of the weighted bootstrap imputation method can be seen. However, still all biases are significantly different from zero. The relative bias for $\hat{P}_1.$ is about 1.5% for large Q for regression and nearest neighbour imputation in comparison to about 2% without the adjustment. The bias for $\hat{P}_2.$ is almost unchanged. The results are given in table 6.8. The weighted bootstrap imputation method is also analysed if restrictions of the size of $\lambda(\ln(Y)|W)$ are introduced, i.e. $\lambda(\ln(\hat{y}^*_{jq})|W) = \min\{\lambda(\ln(\hat{y}^*_{jq})|W),\omega\}$ for all $q=1,...,Q$ and $j \in \bar{r}$, where ω =1.5, 5 and 10. Again mild restrictions, such as ω = 5 and 10, do not have a great effect on the bias, whereas for example for ω =1.5 it is possible to reduce the bias. For example, the bias for $\hat{P}_1.$ is reduced from around 2% to about 1.5% for regression or NN imputation if Q is large. The bias for $\hat{P}_2.$ is slightly increased, however. The results are presented in table 6.9.

Imputation Method	Bias of $\hat{P}_1.$	Rel. Bias $\hat{P}_1.$	Bias of $\hat{P}_2.$	Rel. Bias $\hat{P}_2.$	$V(\hat{P}_1.)$	$MSE(\hat{P}_1.)$
Wboot- reg. imp.(1) Q=20, ω =1/10	$7.54*10^{-4}$ $(0.93*10^{-4})^*$	1.53%	$10.31*10^{-4}$ $(1.54*10^{-4})^*$	0.57%	$8.75*10^{-6}$	$9.32*10^{-6}$
Wboot- NN(1) Q=20, ω =1/10	$7.48*10^{-4}$ $(1.00*10^{-4})^*$	1.52%	$9.81*10^{-4}$ $(1.75*10^{-4})^*$	0.54%	$10.12*10^{-6}$	$10.68*10^{-6}$
Wboot- HDIwr(1) Q=20, ω =1/10	$12.45*10^{-4}$ $(1.05*10^{-4})^*$	2.54%	$23.64*10^{-4}$ $(1.82*10^{-4})^*$	1.30%	$11.02*10^{-6}$	$12.57*10^{-6}$

Table 6.8: Results for the weighted bootsrap imputation method if the denominator of $\lambda(\ln(Y)|W)$ is restricted to a minimum value ω, where ω =1/10.

Imputation Method	Bias of $\hat{P}_1$.	Rel. Bias $\hat{P}_1$.	Bias of $\hat{P}_2$.	Rel. Bias $\hat{P}_2$.	$V(\hat{P}_1.)$	$MSE(\hat{P}_1.)$
Wboot- reg. imp.(1) Q=20, ω =1.5	$7.23*10^{-4}$ $(0.90*10^{-4})^*$	1.47%	$16.09*10^{-4}$ $(1.51*10^{-4})^*$	0.89%	$8.17*10^{-6}$	$8.69*10^{-6}$
Wboot- NN Q=20, ω =1.5	$7.25*10^{-4}$ $(0.97*10^{-4})^*$	1.47%	$15.73*10^{-4}$ $(1.73*10^{-4})^*$	0.54%	$9.55*10^{-6}$	$10.08*10^{-6}$
Wboot- HDIwr Q=20, ω =1.5	$13.16*10^{-4}$ $(1.01*10^{-4})^*$	2.67%	$30.30*10^{-4}$ $(1.79*10^{-4})^*$	1.67%	$10.30*10^{-6}$	$12.02*10^{-6}$

Table 6.9: Results for the weighted bootstrap imputation method where the size of $\lambda(\ln(Y)|W)$ is restricted to a maximum value ω, where ω =1.5.

Imputation Method	Bias of $\hat{P}_1$.	Rel. Bias $\hat{P}_1$.	Bias of $\hat{P}_2$.	Rel. Bias $\hat{P}_2$.	$V(\hat{P}_1.)$	$MSE(\hat{P}_1.)$
Wboot- reg. imp.(1) Q=6	$-7.19*10^{-4}$ $(0.91*10^{-4})^*$	-1.46%	$18.37*10^{-4}$ $(2.16*10^{-4})^*$	1.01%	$8.31*10^{-6}$	$8.83*10^{-6}$
Wboot- NN(1) Q=6	$-7.16*10^{-4}$ $(0.98*10^{-4})^*$	-1.46%	$18.01*10^{-4}$ $(2.38*10^{-4})^*$	0.99%	$9.77*10^{-6}$	$10.02*10^{-6}$
Wboot- HDIwr(1) Q=6	$-2.16*10^{-4}$ $(1.01*10^{-4})^*$	-0.44%	$32.88*10^{-4}$ $(2.44*10^{-4})^*$	1.82%	$10.21*10^{-6}$	$10.25*10^{-6}$

Table 6.10: Results for the weighted bootstrap imputation method reducing the occurrence of large values for $\lambda(\ln(Y)|W)$ by deleting $\ln(y_{jq^*})$ for which $\lambda(\ln(Y)|W)$ is largest, which avoids values $\ln(y_{jq^*})$ that are 'far away' from the minimum, in case of $\sigma_0 > \sigma_1$.

For both adjustment methods we observe a reduction in variance. This is probably due to the smoothing of weights under the adjustment leading to more homogeneous values of $\lambda(\ln(Y)|W)$. Note that both adjustment methods give very similar results. In addition, the adjustment method that excludes $\ln(\hat{y}^*_{jq^*})$ from the set of possible donor values, $\ln(\hat{y}^*_{j1}),...,\ln(\hat{y}^*_{jQ})$, for which the value of $\lambda(\ln(Y)|W)$ is unreasonably large, is applied. The results are given in table 6.10 suggesting a reduction in the bias for $\hat{P}_1$. in comparison to using no adjustment. The bias for $\hat{P}_2$. is slightly increased. The biases are all negative for $\hat{P}_1$.. In particular, we observe a reduction in the

bias for wboot-HDIwr, which would be expected since we now adjust for unusual values of $\ln(\hat{y}^*_{jq})$. We also deleted the two values $\ln(\hat{y}^*_{jq^*})$ for which the values of $\lambda(\ln(Y)|W)$ are largest. However, the results were found to be very similar to the results in table 6.10 and are not reported here.

Despite the bias reduction observed for all three adjustment methods, we can still see a slight bias of the point estimators, which might be caused by the fact that the parameters Φ_0 and σ_0^2 might contain a small bias. Further analysis showed that most coefficients in Φ_0 had a relative bias between 0% and just over 2%, one coefficient was biased by 9%. When these parameters were specified correctly by fitting $\hat{f}(\ln(Y)|W, I=0)$ directly, which is only possible in a simulation study, the bias of the weighted bootstrap imputation method tended to zero if Q was large. This implies that misspecifications of the parameters might be the cause for the bias observed here. Note that an analysis of the residuals from the models used in the imputation process showed that the residuals were approximately normally distributed. The assumption of normality as made in B therefore appears to be reasonable in the simulation study.

b.2) Performance of the Weighted Bootstrap Imputation Method under Misspecification

Here, the weighted bootstrap imputation method is analysed under different misspecifications to test the robustness of the method. First, we allow the covariates employed in the model to generate $\ln(X)$, the model to generate $\ln(Y)$, the nonresponse model and the covariates W employed in the imputation models to differ[2]. In particular, the imputation model only contains a

[2] The model generating *ln(X)* contains:
major occupation group (SOC), part-time (PT), Age and Age squared, qualification (Q), length of time continuously employed (EMPMON), less than weekly (LTWK), head of household (HOH), married, number of employees at work place (SIZE), gender (FEMALE) and interactions between SOC and PT. (=model in appendix A3.1)
The model generating *ln(Y)* contains:
All the variables specified in table 2.1 in chapter 2.
The model generating nonresponse contains:
All the variables specified in appendix 3.2 (Model A3)
The covariates, denoted *W*, employed in imputation models contain:
major occupation group (SOC), part-time (PT), qualification (Q), Age, length of time continuously employed (EMPMON),married (MARRIED), number of employees at work place (SIZE), industry (IND), region (REG), gender (FEMALE).

subset of the variables employed in the model generating Y. It does not include interactions and squared terms. The performances of the weighted bootstrap method as well as the derived imputation method under these conditions are given in table 6.11. The bias is now negative which is probably caused by the fact that in this particular simulation study we had mostly $\hat{\sigma}_0 < \hat{\sigma}_1$, which seemed to be due to the particular choice of variables. The bias is less than 1%, however, still significant for some cases. We can conclude that under the misspecifications considered here the weighted bootstrap method still seems to perform well.

Imputation Method	Bias of $\hat{P}_1$.	Rel. Bias $\hat{P}_1$.	Bias of $\hat{P}_2$.	Rel. Bias $\hat{P}_2$.	$V(\hat{P}_1.)$	$MSE(\hat{P}_1.)$
Wboot- reg. imp.(1) Q=6	$-1.13*10^{-4}$ $(0.95*10^{-4})$	-0.24 %	$-6.44*10^{-4}$ $(1.64*10^{-4})^*$	-0.35 %	$9.20*10^{-6}$	$9.21*10^{-6}$
Wboot- reg. imp.(1) Q=10	$-1.47*10^{-4}$ $(0.97*10^{-4})$	-0.31 %	$-11.94*10^{-4}$ $(1.70*10^{-4})^*$	-0.66 %	$9.48*10^{-6}$	$9.50*10^{-6}$
Wboot- NN(1) Q=6	$-3.87*10^{-4}$ $(0.99*10^{-4})^*$	-0.83 %	$1.47*10^{-4}$ $(1.83*10^{-4})$	0.08 %	$9.82*10^{-6}$	$9.97*10^{-6}$
Wboot- NN(1) Q=10	$-4.35*10^{-4}$ $(0.99*10^{-4})^*$	-0.93 %	$-5.40*10^{-4}$ $(1.86*10^{-4})^*$	-0.29 %	$9.97*10^{-6}$	$10.16*10^{-6}$
Wboot- HDIwr(1) Q=6	$-0.70*10^{-4}$ $(1.03*10^{-4})$	-0.16 %	$12.59*10^{-4}$ $(1.87*10^{-4})^*$	0.69 %	$10.75*10^{-6}$	$10.75*10^{-6}$
Wboot- HDIwr(1) Q=10	$-1.21*10^{-4}$ $(1.04*10^{-4})$	-0.25 %	$5.61*10^{-4}$ $(1.90*10^{-4})^*$	0.31 %	$10.88*10^{-6}$	$10.89*10^{-6}$

Table 6.11: Performance of the weighted bootstrap imputation method under misspecification of the covariates W.

The performance of the weighted bootstrap method is also investigated under the assumption of MAR, instead of the assumption of CME, on which the method is theoretically based. The data is generated as described at the beginning of this section b.2). The method shows a slightly higher bias than under the correct assumption of CME, but still performs well with a relative bias for both point estimators of 1% or less for wboot-NN and wboot-regression imputation. The results are given in table 6.12. The good performance of the weighted bootstrap imputation under this misspecification is a potential attraction of the method.

The case where $\ln(X)$ is not generated but bootstrapped from LFS data is also analysed. The advantage is that this simulation study reflects the distribution of $\ln(X)$ more realistically and therefore features the characteristics of the data in the LFS better, although some units may be duplicated in $s^{(a)}$. The results are given in table 6.13. We can see that even under these mild misspecifications all methods are significantly biased with about 5% relative bias for $\hat{P}_1$. using wboot-NN and wboot-reg. imp. The wboot-HDIwr and the derived imputation method show the largest bias for $\hat{P}_1$. with 15% and 26% respectively.

Imputation Method	Bias of $\hat{P}_1$.	Rel. Bias $\hat{P}_1$.	Bias of $\hat{P}_2$.	Rel. Bias $\hat{P}_2$.	$V(\hat{P}_1.)$	$MSE(\hat{P}_1.)$
Wboot- reg. imp.(1) Q=6	$5.07*10^{-4}$ $(0.93*10^{-4})^*$	1.03 %	$2.54*10^{-4}$ $(1.46*10^{-4})$	0.14 %	$8.70*10^{-6}$	$8.96*10^{-6}$
Wboot- NN(1) Q=6	$4.96*10^{-4}$ $(1.01*10^{-4})^*$	1.01 %	$2.53*10^{-4}$ $(1.68*10^{-4})$	0.14 %	$10.58*10^{-6}$	$10.58*10^{-6}$
Wboot- HDIwr(1) Q=6	$13.47*10^{-4}$ $(1.07*10^{-4})^*$	2.74 %	$22.88*10^{-4}$ $(1.79*10^{-4})^*$	1.26 %	$11.46*10^{-6}$	$13.28*10^{-6}$

Table 6.12: Results for the weighted bootstrap imputation method with Q=6, based on regression imputation, nearest neighbour and hot deck imputation, under the assumption of MAR nonresponse.

Imputation Method	Bias of $\hat{P}_1$.	Rel. Bias $\hat{P}_1$.	Bias of $\hat{P}_2$.	Rel. Bias $\hat{P}_2$.	$V(\hat{P}_1.)$	$MSE(\hat{P}_1.)$
Wboot- reg. imp.(1) Q=6	$28.38*10^{-4}$ $(0.91*10^{-4})^*$	5.07 %	$42.26*10^{-4}$ $(1.47*10^{-4})^*$	2.27 %	$8.34*10^{-6}$	$1.64*10^{-6}$
Wboot- NN(1) Q=6	$27.50*10^{-4}$ $(1.00*10^{-4})^*$	4.91 %	$42.77*10^{-4}$ $(1.73*10^{-4})^*$	2.30 %	$10.05*10^{-6}$	$17.62*10^{-6}$
Wboot- HDIwr(1) Q=6	$78.63*10^{-4}$ $(1.12*10^{-4})^*$	15.08 %	$91.85*10^{-4}$ $(1.78*10^{-4})^*$	4.64 %	$12.63*10^{-6}$	$74.47*10^{-6}$
Derived imputation(1)	$139.7*10^{-4}$ $(1.44*10^{-4})^*$	26.82 %	$-17.23*10^{-4}$ $(1.40*10^{-4})^*$	-0.87 %	$20.82*10^{-6}$	$216.2*10^{-6}$

Table 6.13: Results for the weighted bootstrap imputation method and the derived imputation method where $\ln(X)$ is not generated.

Analysing the bias of the estimators of the parameters Φ_0 and σ_0, in the case of not generating $\ln(X)$ it was found that the estimators are biased when compared to the true estimates obtained directly from $f(\ln(Y)|W,I=0)$. The relative biases for most of the coefficients are approximately between 5% and 40% with some very few coefficients showing an even higher relative bias. This indicates that the formulae deriving estimates of Φ_0 and σ_0 as specified in equations (6.34) and (6.35) are potentially problematic. It was tried to improve the estimation of Φ_0 and σ_0 by excluding units from the estimation of the models, that are involved in deriving estimates for Φ_0 and σ_0, where the standardized residuals exceed a certain value. However, this did not show a significant improvement in the estimation of Φ_0 and σ_0. An analysis of the residuals of all models in the imputation process showed that the residuals seemed to be approximately normally distributed. The assumption of normality as made in B therefore appears to be reasonable. It seems that the misspecification of the parameters using the derivations above might be the reason for the unsatisfactory performance of the weighted bootstrap imputation method under certain conditions.

Since the weighted bootstrap imputation method leads to biased results if the variable $\ln(X)$ is not generated adjustment methods are applied, restricting the denominator of $\lambda(\ln(Y)|W)$ or the size of $\lambda(\ln(Y)|W)$. The results are given in table 6.14 and table 6.15 respectively. However, even with these adjustments we still observe a relative bias of about 4% for $\hat{P}_1$.. Further investigation revealed the occurrence of unusually large values of $\lambda(\ln(Y)|W)$ even after adjustment. We therefore apply the third adjustment method, as described in section 6.3.4.2, that excludes $\ln(\hat{y}^*_{jq^*})$ from the set of possible donor values, $\ln(\hat{y}^*_{j1}),...,\ln(\hat{y}^*_{jQ})$, for which the value of $\lambda(\ln(Y)|W)$ is unreasonably large. The results are given in table 6.16. The relative bias for the estimator $\hat{P}_1$. is reduced to less than 1%. Note that the bias of $\hat{P}_2$. is still just above 2%. The performance of the weighted bootstrap imputation method under stronger misspecification of models and assumptions involved are not analysed here. Overall, the weighted bootstrap imputation method seems sensitive even to mild misspecification and requires the use of adjustment methods.

Imputation Method	Bias of $\hat{P}_1.$	Rel. Bias $\hat{P}_1.$	Bias of $\hat{P}_2.$	Rel. Bias $\hat{P}_2.$	$V(\hat{P}_1.)$	$MSE(\hat{P}_1.)$
Wboot- reg. imp.(1) Q=6, ω =1/10	$22.32*10^{-4}$ $(0.89*10^{-4})^*$	3.99%	$43.23*10^{-4}$ $(1.45*10^{-4})^*$	2.32%	$7.92*10^{-6}$	$12.91*10^{-6}$
Wboot- NN(1) Q=6, ω =1/10	$21.64*10^{-4}$ $(0.97*10^{-4})^*$	3.86%	$43.70*10^{-4}$ $(1.73*10^{-4})^*$	2.35%	$9.58*10^{-6}$	$14.26*10^{-6}$
Wboot- HDIwr(1) Q=6, ω =1/10	$24.19*10^{-4}$ $(1.01*10^{-4})^*$	4.32%	$63.69*10^{-4}$ $(1.73*10^{-4})^*$	3.42%	$10.24*10^{-6}$	$16.09*10^{-6}$

Table 6.14: Results for the weighted bootstrap imputation method, where $\ln(X)$ is not generated, restricting the denominator of $\lambda(\ln(Y)|W)$.

Imputation Method	Bias of $\hat{P}_1.$	Rel. Bias $\hat{P}_1.$	Bias of $\hat{P}_2.$	Rel. Bias $\hat{P}_2.$	$V(\hat{P}_1.)$	$MSE(\hat{P}_1.)$
Wboot- reg. imp.(1) Q=6, ω =1.5	$22.94*10^{-4}$ $(0.87*10^{-4})^*$	4.10%	$44.45*10^{-4}$ $(1.45*10^{-4})^*$	2.39%	$7.63*10^{-6}$	$12.89*10^{-6}$
Wboot- NN(1) Q=6, ω =1.5	$22.19*10^{-4}$ $(0.96*10^{-4})^*$	3.96%	$44.77*10^{-4}$ $(1.73*10^{-4})^*$	2.40%	$9.24*10^{-6}$	$14.16*10^{-6}$
Wboot- HDIwr(1) Q=6, ω =1.5	$26.36*10^{-4}$ $(0.99*10^{-4})^*$	4.71%	$64.82*10^{-4}$ $(1.74*10^{-4})^*$	3.47%	$9.90*10^{-6}$	$16.85*10^{-6}$

Table 6.15: Results for the weighted bootstrap imputation method, where $\ln(X)$ is not generated, restricting the size of $\lambda(\ln(Y)|W)$.

Imputation Method	Bias of $\hat{P}_1.$	Rel. Bias $\hat{P}_1.$	Bias of $\hat{P}_2.$	Rel. Bias $\hat{P}_2.$	$V(\hat{P}_1.)$	$MSE(\hat{P}_1.)$
Wboot- reg. imp.(1) Q=6, L=1	$-1.07*10^{-4}$ $(0.84*10^{-4})^*$	-0.19 %	$44.06*10^{-4}$ $(1.80*10^{-4})^*$	2.37 %	$7.11*10^{-6}$	$7.12*10^{-6}$
Wboot- NN(1) Q=6, L=1	$-1.31*10^{-4}$ $(0.91*10^{-4})^*$	-0.23 %	$43.57*10^{-4}$ $(2.03*10^{-4})^*$	2.34 %	$8.38*10^{-6}$	$8.39*10^{-6}$
Wboot- HDIwr(1) Q=6, L=1	$-0.62*10^{-4}$ $(0.93*10^{-4})^*$	-0.11 %	$64.57*10^{-4}$ $(2.11*10^{-4})^*$	3.47 %	$8.80*10^{-6}$	$8.80*10^{-6}$

Table 6.16: Results for the weighted bootstrap imputation method, where $\ln(X)$ is not generated, adjusting against large values of $\lambda(\ln(Y)|W)$ by deleting $\ln(y_{jq^*})$ for which $\lambda(\ln(Y)|W)$ is largest. L=1 indicates that only one value $\ln(y_{jq^*})$ is deleted.

6.3.4.4 Application of the Weighted Bootstrap Imputation Method to LFS Data

The two hot deck weighted bootstrap imputation methods, wboot- HDIwr(1) and wboot- NN(1), are applied to LFS data. Regression imputation was not implemented since it does not have desirable characteristics when applied to earnings data, as discussed in section 2.2.2. The quarter March-May 2000 is used. For comparison we also apply PSW, NN(10) and HDIwr(10), as derived under the assumption of MAR, and the derived imputation method valid under the CME assumption. The results are presented in table 6.17. Estimates of proportions of employees that are less well paid obtained for a method valid under the CME assumption may be expected to be lower than for a method valid under the MAR assumption. This is because under MAR it is assumed that the conditional distribution of Y for the respondents is the same as for the nonrespondents. The nonresponse is independent of true hourly pay given information on other covariates. However, those with higher wages are less likely to be paid by the hour and are therefore less likely to respond. Methods that are valid under the MAR assumption may lead to an overestimate of proportions of employees earning at the lower end of the pay scale, whereas the CME assumption allows for the fact that the nonresponse depends on true hourly pay. This

argument is also discussed in Manning and Dickens (2002, p. 29). Empirical evidence of this hypothesis can be found in the performance of the MAR-based methods under the CME assumption in section 6.2.2 showing a positive bias.

Imputation Method	$\hat{P}_1$. (weighted, for 18+)	$\hat{P}_2$. (weighted, for 18+)
PSW under MAR	0.54%	27.10%
NN(10) under MAR	0.55%	26.61%
HDIwr(10) under MAR	0.57%	26.01%
Wboot- NN(1), Q=6	0.99%	25.99%
Wboot- NN(1), Q=6, ω =1/100	0.59%	25.50%
Wboot- NN(1), Q=6, ω =1/10	0.53%	25.46%
Wboot- NN(1), Q=6, ω =10	0.43%	25.82%
Wboot- NN(1), Q=6, ω =5	0.46%	26.02%
Wboot- NN(1), Q=6, ω =1.5	0.42%	25.91%
Wboot- NN(1), Q=6, deleting ln(y_{jq^*}) where λ large, L=1	0.39%	25.96%
Wboot- NN(1), Q=6, deleting ln(y_{jq^*}) where λ large, L=2	0.38%	25.17%
Wboot- HDIwr, Q=6	0.94%	26.54%
Wboot- HDIwr, Q=6, ω =1/100	0.84%	26.80%
Wboot- HDIwr, Q=6, ω =1/10	0.84%	26.45%
Wboot- HDIwr, Q=6, ω =10	0.82%	26.22%
Wboot- HDIwr, Q=6, ω =5	0.76%	26.74%
Wboot- HDIwr, Q=6, ω =1.5	0.63%	26.45%
Wboot- HDIwr, Q=6, deleting ln(y_{jq^*}) where λ large, L=1	0.37%	26.86%
Wboot- HDIwr, Q=6, deleting ln(y_{jq^*}) where λ large, L=2	0.39%	26.92%
Derived imputation method	2.63%	23.30%

Table 6.17: Estimates for P_1 and P_2 based on wboot-HDIwr and wboot-NN. For comparison the results for PSW, NN(10) and HDIwr(10) under MAR as well as the derived imputation are presented, March-May 2000 LFS quarter.

However, the weighted bootstrap imputation method without adjustment seems to indicate an overestimation of the proportion of low paid employees. Adjustments against a small denominator as well as adjustments against large values of $\lambda(\ln(Y)|W)$ reduce the estimates of the proportion of low paid employees. Adjustments against large values of $\lambda(\ln(Y)|W)$ seem to have a greater impact on the reduction of the estimates than adjustment for the denominator. Only in the case of wboot-NN with adjustment against large values of $\lambda(\ln(Y)|W)$, we have estimates that are below the estimates from the MAR-based methods. The wboot-HDI method seems to overestimate $\hat{P}_.$ even with adjustment against large values of $\lambda(\ln(Y)|W)$ and against a small denominator. This seems to indicate that these two adjustment methods might not be fully appropriate and that further adjustment avoiding large values of $\lambda(\ln(Y)|W)$ may need to be applied. We therefore exclude values $\ln(y_{jq^*})$ that have the furthest distance from the minimum $-c_3/2c_2$. Under this adjustment the estimate of $\hat{P}_1.$ is smaller than the estimate from the methods under the MAR assumption. Overall, wboot-NN is preferred to wboot-HDI. However, the weighted bootstrap method seems to give very inconsistent estimates of the two parameters of interest depending if and which adjustment method is used.

6.3.4.5 Conclusions on Weighted Bootstrap Imputation Method

The weighted bootstrap imputation method seems to perform well under ideal conditions. However, it is advisable to restrict either the denominator of $\lambda(\ln(Y)|W)$ or the size of $\lambda(\ln(Y)|W)$ to avoid the occurrence of unreasonably large values of $\lambda(\ln(Y)|W)$ even under ideal conditions. These adjustment methods lead to a reduction in the variance of the estimator, probably because of the smoothing of weights. The use of the weighted bootstrap fractional imputation also leads to a gain in efficiency using effectively weighted repeated imputation. The results of the weighted bootstrap method depend on the choice of Q. Generally, the larger Q the smaller is the bias.

Disadvantages of the weighted bootstrap method are the occurrence of some large values of $\lambda(\ln(Y)|W)$, especially under misspecification, and the fact that $\hat{p}^*_{jq}$ strongly depends on the specific values $\ln(\hat{y}^*_{j1}),...,\ln(\hat{y}^*_{jQ})$ selected as potential donor values for $j \in r$. If already one value $\ln(\hat{y}^*_{jq^*})$ is 'very different' from the other selected values for imputation this value might be

allocated a large probability of selection. One reason for the occurrence of large values for $\lambda(\ln(Y)|W)$ might be misspecifications of Φ_0 and σ_0^2 using the derivations (6.34) and (6.35). Analysis has shown that under model misspecification it is difficult to obtain approximately unbiased estimates of Φ_0 and σ_0^2, which has an effect on the form of $\lambda(\ln(Y)|W)$. The form of $\lambda(\ln(Y)|W)$ strongly depends on the specifications of the parameters Φ_0 and σ_0^2, which is generally a robustness concern. An investigation of the residuals of the models involved in the imputation process indicates that the assumption of normality as made in assumption B seems plausible. The application of adjustment methods, such as restricting the denominator of $\lambda(\ln(Y)|W)$ or defining an upper bound of $\lambda(\ln(Y)|W)$, improves the performance of the weighted bootstrap method. However, under misspecification also these adjustments do not seem to reduce the bias much. Due to these drawbacks the variance and variance estimation of the weighted Bootstrap method is not considered here.

In summary, the strong dependence of the weighted bootstrap method on adjustments based on more or less arbitrarily chosen thresholds is a potential drawback of this imputation method. For example, the method can provide different estimates under different thresholds for adjustment. The method wboot-HDIwr within classes seems to be more prone to error since the defined imputation classes cover a wider range of values which can cause the occurrence of unusually large values of $\lambda(\ln(Y)|W)$. Overall, weighted bootstrap imputation based on nearest neighbour with adjustment against large values of $\lambda(\ln(Y)|W)$, preferably using weighting, is preferred. The derived imputation method seems to be sensitive to model misspecification and relies strongly on the correct specifications of the parameters Φ_0 and σ_0^2. It is therefore not recommended for practical use.

6.3.5 Imputation using Data Augmentation Based on the CME Assumption

The following presents an imputation method, which is based on the CME assumption as opposed to the assumption of MAR, using data augmentation in a Bayesian framework. This approach allows us to impute values from the distribution $f(Y \mid X, W, I = 0)$ for the nonrespondents, which is assumed to differ from the distribution of the respondents. The novelty is the use of data augmentation under nonignorable nonresponse based on the CME assumption 6.1. The method is therefore an extension to data augmentation in the context of missing data, which is commonly applied under the assumption of MAR. In addition, the use of hot deck imputation in the imputation step of data augmentation both under MAR and under CME instead of the traditional parametric random regression imputation is proposed. This approach has the advantage that actually observed values are used for imputation and that specific features such as truncation and steps in the distribution of interest can be preserved, which is of importance in the context of earnings distributions. The use of predictive mean matching imputation within data augmentation also allows for non-constant variance of the residuals. Another advantage of using hot deck imputation is that the resulting estimator can be expressed as a weighted estimator as discussed in chapter 5.

The section is structured as follows. Section 6.3.5.1 gives a brief introduction to Markov chain Monte Carlo methods with particular emphasise on Gibbs sampling and data augmentation. The application of data augmentation in the missing data context under the assumption of MAR is described in section 6.3.5.2. In section 6.3.5.3 data augmentation under the common measurement error model is derived making use of rejection sampling and the weighted bootstrap method. In section 6.3.5.4 a simulation study is carried out investigating the properties of the method under ideal conditions following the assumptions made and under misspecification of some of the underlying assumptions. In addition to parametric random regression imputation two forms of hot deck imputation are considered in the data augmentation procedure. Finally, in section 6.3.5.5 the proposed method is applied to the LFS to obtain estimates of low pay under the assumption of CME rather than MAR. The results are compared to estimates derived under single and multiple imputation methods valid under the assumption of MAR. Some concluding remarks are given in section 6.3.5.6.

6.3.5.1 Review of Gibbs Sampling and Data Augmentation

Bayesian analysis requires specifying a likelihood as the conditional density of the data given the parameters and a prior distribution representing knowledge about the parameters prior to data collection. The product of the prior and likelihood is the joint density of the data and the parameters. The posterior density, dividing the joint density by the marginal density of the data, is the basis for all Bayesian inference and summarizes all the information about the parameters of interest. Computing the posterior, however, is often difficult and requires high-dimensional numerical integration. This computational problem can be addressed by methods using Markov chain Monte Carlo integration, such as Gibbs sampling, which generates a Markov chain whose stationary distribution is the posterior distribution of interest. Closely related to Gibbs sampling is the method of data augmentation, which is widely used in the missing data context. Under certain conditions data augmentation is a special case of Gibbs sampling. The advantage of such methods is that the desired chain can be simulated using only the joint density of the parameters and the data, i.e. the product of the likelihood and the prior and not the unknown posterior. The observations in a sample from the chain are approximately identically distributed with common distribution to the required posterior. The use of Gibbs sampling and data augmentation are described in detail in Tanner (1996), Gilks, Richardson and Spiegelhalter (1996), Gelman et al. (1998), and with particular reference to missing data in Schafer (1997) and Carroll, Ruppert and Stefanski (1995).

Gibbs sampling and data augmentation are now briefly reviewed. In general, Markov chain Monte Carlo methods describe techniques for pseudorandom draws from probability distributions. The aim of such methods is to generate values of a random variable, here denoted Z. The density of Z, $f(Z)$, is denoted $B(Z)$, which is called the target distribution, such that $f(Z)=B(Z)$. Instead of directly drawing from the distribution B a sequence $\{\tilde{Z}^{(1)},\tilde{Z}^{(2)},...,\tilde{Z}^{(d)},...\}$ is generated, where each variate $\tilde{Z}^{(d)}$ depends on the proceeding values $\tilde{Z}^{(1)},...,\tilde{Z}^{(d-1)}$ and very importantly the stationary distribution, i.e. the limiting marginal distribution of $\tilde{Z}^{(d)}$, when d tends to infinity, is the desired target function B,

$$\lim_{d\to\infty} f(\tilde{Z}^{(d)}) = B(\tilde{Z}). \tag{6.45}$$

If d is sufficiently large $\tilde{Z}^{(d)}$ is an approximately random draw from the distribution B. The advantage of this approach generating a sequence $\tilde{Z}^{(d)}$ rather than directly drawing from B can be seen if B is difficult to draw from directly but it is relatively easy to draw variates out of the sequence. One way of generating such a sequence is the use of Gibbs sampling. Let us assume that a random vector $\tilde{Z}$ is subdivided into k subvectors, such that $\tilde{Z}=(\tilde{Z}_1,\tilde{Z}_2,...,\tilde{Z}_k)$. The distribution $f(\tilde{Z})$ denotes the joint distribution of $\tilde{Z}$, which is the target distribution. The aim is to find a sequence $\tilde{Z}^{(d)}$ with the stationary distribution $f(\tilde{Z})$. To obtain such a sequence Gibbs sampling iteratively draws from the conditional distribution of each subvector given all other vectors. Having obtained the value of $\tilde{Z}^{(d)}$ the value for $\tilde{Z}^{(d+1)}$ is obtained by successively drawing from the following distributions

$$\begin{aligned}
\tilde{Z}_1^{(d+1)} &\sim f(\tilde{Z}_1 \mid \tilde{Z}_2^{(d)},...,\tilde{Z}_k^{(d)}) \\
\tilde{Z}_2^{(d+1)} &\sim f(\tilde{Z}_2 \mid \tilde{Z}_1^{(d+1)},\tilde{Z}_3^{(d)},...,\tilde{Z}_k^{(d)}) \\
&... \\
\tilde{Z}_k^{(d+1)} &\sim f(\tilde{Z}_k \mid \tilde{Z}_1^{(d+1)},\tilde{Z}_3^{(d+1)},...,\tilde{Z}_{k-1}^{(d+1)}).
\end{aligned} \tag{6.46}$$

We therefore draw from the conditional distributions of $\tilde{Z}_1$, $\tilde{Z}_2$, ..., $\tilde{Z}_k$ conditioning each time on the most recently drawn values of all other subvectors. Having obtained the full set for $\tilde{Z}^{(d+1)}$ the values for $\tilde{Z}^{(d+2)}$, $\tilde{Z}^{(d+3)}$ etc. can be obtained in the same way. By applying this approach a Markov chain of the form $\{\tilde{Z}^{(d)},d=0,1,2,...\}$ is obtained, which has the distribution $f(\tilde{Z})$ as the stationary distribution.

Closely related to the Gibbs sampler is data augmentation. Under certain conditions data augmentation is a special case of the Gibbs sampler. Following the notation in Schafer (1997) let $\tilde{z}$ denote a random vector, which is subdivided into two subvectors $\tilde{u}$ and $\tilde{v}$, such that $\tilde{z}=(\tilde{u},\tilde{v})$. Let us assume that the joint distribution $f(\tilde{z})=B(\tilde{z})$ cannot easily be simulated but the conditional distributions $f(\tilde{u}\mid\tilde{v})=g(\tilde{u}\mid\tilde{v})$ and $f(\tilde{v}\mid\tilde{u})=h(\tilde{v}\mid\tilde{u})$ can. The aim is to find a sequence $\tilde{z}^{(d)}=(\tilde{u}^{(d)},\tilde{v}^{(d)})$, whose distribution approximates the joint distribution $f(\tilde{z})$, if d is large enough such that $\lim_{d\to\infty} f(\tilde{z}^{(d)})=B(\tilde{z})$. Let

$$\tilde{Z}^{(d)}=(\tilde{z}_1^{(d)},\tilde{z}_2^{(d)},...,\tilde{z}_q^{(d)})=((\tilde{u}_1^{(d)},\tilde{v}_1^{(d)}),(\tilde{u}_2^{(d)},\tilde{v}_2^{(d)}),...,(\tilde{u}_q^{(d)},\tilde{v}_q^{(d)})),$$

such that $\tilde{Z}^{(d)}$ is a sample of size q from a distribution that approximates the target distribution $f(\tilde{z})$ at iteration d. To obtain the values for

$$\tilde{Z}^{(d+1)} = (\tilde{z}_1^{(d+1)}, \tilde{z}_2^{(d+1)}, ..., \tilde{z}_q^{(d+1)}) = ((\tilde{u}_1^{(d+1)}, \tilde{v}_1^{(d+1)}), (\tilde{u}_2^{(d+1)}, \tilde{v}_2^{(d+1)}), ..., (\tilde{u}_q^{(d+1)}, \tilde{v}_q^{(d+1)}))$$

we first obtain $$\tilde{U}^{(d+1)} = (\tilde{u}_1^{(d+1)}, \tilde{u}_2^{(d+1)}, ..., \tilde{u}_q^{(d+1)})$$

by drawing $$\tilde{u}_i^{(d+1)} \sim g(\tilde{u} \mid \tilde{v}_i^{(d)}) = f(\tilde{u} \mid \tilde{v}_i^{(d)}) \text{ for all } i = 1, ..., q \tag{6.47}$$

independently. Secondly, we obtain

$$\tilde{V}^{(d+1)} = (\tilde{v}_1^{(d+1)}, \tilde{v}_2^{(d+1)}, ..., \tilde{v}_q^{(d+1)})$$

by drawing independently

$$\tilde{v}_i^{(d+1)} \sim h(\tilde{v} \mid \tilde{u}_i^{(d+1)}) = f(\tilde{v} \mid \tilde{u}_i^{(d+1)}) \text{ for all } i = 1, ..., q. \tag{6.48}$$

As shown in Tanner and Wong (1982) the distribution $\tilde{z}^{(d)}$ converges to $f(\tilde{z}) = B(\tilde{z})$ if d tends to infinity. The algorithm as described here represents q parallel runs of the Gibbs sampler. For the case of q=1 data augmentation is therefore a special case of the Gibbs sampler. An alternative way of drawing $\tilde{V}^{(d+1)}$ as described in Tanner and Wong (1982) is to draw the required values as an iid sample from the equally weighted mixture of the conditionals $h(\tilde{v} \mid \tilde{u}_i^{(d+1)})$, i.e.

$$\bar{h}(\tilde{v} \mid \tilde{U}^{(d+1)}) = \frac{1}{q} \sum_{i=1}^{q} h(\tilde{v} \mid \tilde{u}_i^{(d+1)}). \tag{6.49}$$

When q is large this provides a good approximation to the marginal density $f(\tilde{v}) = \int f(\tilde{u}, \tilde{v}) d\tilde{u}$.

6.3.5.2 Application of Data Augmentation to Missing Data Under the Assumption of MAR

Of particular interest is the application of data augmentation to incomplete data, which is described in detail in Schafer (1997). The general case is briefly discussed where $\tilde{H}$ denotes the

complete data-matrix, partitioned in $\tilde{H}_{obs}$ and $\tilde{H}_{mis}$, i.e. observed and missing data respectively, $\tilde{I}$ is an indicator matrix indicating if person i responded to item j. The parameter ς refers to the parameter of the model for the complete data $\tilde{H}$ and ψ refers to the parameter of the missingness mechanism, using the notation as introduced in chapter 1. Initially, MAR is assumed such that ψ does not need to be considered. The algorithm of data augmentation in the context of missing data may be motivated by the following two expressions. The posterior density of the missing data can be written as

$$f(\tilde{H}_{mis} \mid \tilde{H}_{obs}) = \int f(\tilde{H}_{mis} \mid \tilde{H}_{obs}, \varsigma) f(\varsigma \mid \tilde{H}_{obs}) d\varsigma , \tag{6.50}$$

the conditional predictive distribution of $\tilde{H}_{mis}$ given ς, averaged over the observed-data posterior of ς, $f(\varsigma \mid \tilde{H}_{obs})$. On the other hand the desired observed-data posterior distribution can be written as

$$f(\varsigma \mid \tilde{H}_{obs}) = \int f(\varsigma \mid \tilde{H}_{obs}, \tilde{H}_{mis}) f(\tilde{H}_{mis} \mid \tilde{H}_{obs}) d\tilde{H}_{mis} . \tag{6.51}$$

The integral in (6.50) contains the density $f(\varsigma \mid \tilde{H}_{obs})$ and in return the integral in (6.51) contains the expression $f(\tilde{H}_{mis} \mid \tilde{H}_{obs})$. This dependency between $f(\tilde{H}_{mis} \mid \tilde{H}_{obs})$ and $f(\varsigma \mid \tilde{H}_{obs})$ motivates an iterative algorithm to calculate $f(\varsigma \mid \tilde{H}_{obs})$. In many cases the integrals in (6.50) and (6.51) are difficult to calculate analytically. However, it is often possible to perform the integration iteratively using methods such as data augmentation. Given a current estimate of $\tilde{H}_{mis}$ it is necessary that we can calculate the distribution $f(\varsigma \mid \tilde{H}_{obs}, \tilde{H}_{mis})$ or sample from it numerically. Note that $f(\tilde{H}_{mis} \mid \tilde{H}_{obs})$ does not rely on the observed response pattern I under the assumption of MAR.

Often the observed-data posterior $f(\varsigma \mid \tilde{H}_{obs})$ is intractable and cannot easily be summarized or simulated. The idea of data augmentation is to augment $\tilde{H}_{obs}$ by the quantity $\tilde{H}_{mis}$. This procedure will be referred to as imputing the values for $\tilde{H}_{mis}$. It is assumed that the complete-data posterior $f(\varsigma \mid \tilde{H}_{obs}, \tilde{H}_{mis})$, also called the augmented data posterior, can be calculated. However, the posterior density of interest is $f(\varsigma \mid \tilde{H}_{obs})$. A way of finding an approximation to this distribution is to generate a sequence with the stationary distribution $f(\varsigma \mid \tilde{H}_{obs})$. This procedure also raises questions of how to impute appropriate values for $\tilde{H}_{mis}$. We will therefore apply the following iterative sampling scheme corresponding to data augmentation with q=1 to approximate

$f(\varsigma \mid \tilde{H}_{obs})$ and to find appropriate imputed values for $\tilde{H}_{mis}$. Using the previous notation we have for the special case of data augmentation with q=1

$$\tilde{Z}^{(d)} = \{\tilde{z}_1^{(d)}\} = \{\tilde{u}^{(d)}, \tilde{v}^{(d)}\} = \{\tilde{H}_{mis}^{(d)}, \varsigma^{(d)}\}.$$

Data augmentation is divided into the following two steps. Given a current estimate of the parameter, $\varsigma^{(d)}$, values for $\tilde{H}_{mis}^{(d+1)}$ are obtained by

$$\tilde{H}_{mis}^{(d+1)} = \tilde{u}^{(d+1)} \sim g(\tilde{u} \mid \tilde{v}^{(d)}) = f(\tilde{u} \mid \tilde{v}^{(d)}) = f(\tilde{H}_{mis} \mid \tilde{H}_{obs}, \varsigma^{(d)}),$$

which means that the imputed values for the missing data are obtained by drawing from the conditional predictive distribution of $\tilde{H}_{mis}$:

$$\tilde{H}_{mis}^{(d+1)} \sim f(\tilde{H}_{mis} \mid \tilde{H}_{obs}, \varsigma^{(d)}). \tag{6.52}$$

This step is called the *I*-step (*Imputation step*) and describes the imputation of $\tilde{H}_{mis}$. Having obtained $\tilde{H}_{mis}^{(d+1)}$ the values for $\varsigma^{(d+1)}$ are drawn from the complete-data posterior of the parameter ς:

$$\varsigma^{(d+1)} = \tilde{v}^{(d+1)} \sim h(\tilde{v} \mid \tilde{u}^{(d+1)}) = f(\tilde{v} \mid \tilde{u}^{(d+1)}) = f(\varsigma \mid \tilde{H}_{obs}, \tilde{H}_{mis}^{(d+1)}) \tag{6.53}$$

using the second step (6.48) in the data augmentation procedure. This step is referred to as the *P*-step (*Posterior step*) and describes drawing a value for ς from the complete-data posterior. The *I*-step and the *P*-step are motivated by the above expressions (6.50) and (6.51) respectively (Tanner and Wong, 1982). Starting with an initial value of the parameters $\varsigma^{(0)}$ and repeating the two steps a sequence $\{\tilde{Z}^{(d)}, d = 1,2,...\} = \{(\tilde{H}_{mis}^{(d)}, \varsigma^{(d)}), d = 1,2,...\}$ is obtained, which has the stationary distribution $f(\varsigma, \tilde{H}_{mis} \mid \tilde{H}_{obs})$. More importantly the subsequence $\{\varsigma^{(d)}, d = 1,2,...\}$ has the stationary distribution $f(\varsigma \mid \tilde{H}_{obs})$ and the subsequence $\{\tilde{H}_{mis}^{(d)}, d = 1,2,...\}$ converges to $f(\tilde{H}_{mis} \mid \tilde{H}_{obs})$. We can therefore regard $\varsigma^{(d)}$ as an approximate draw from $f(\varsigma \mid \tilde{H}_{obs})$ and similarly $\tilde{H}_{mis}^{(d)}$ as an approximate draw from $f(\tilde{H}_{mis} \mid \tilde{H}_{obs})$ if d is large enough. To implement the algorithm it is required that we are able to sample from the two distributions $f(\tilde{H}_{mis} \mid \tilde{H}_{obs}, \varsigma)$ and $f(\varsigma \mid \tilde{H}_{obs}, \tilde{H}_{mis})$. It is possible to use a single chain with $q = 1$ and $d = 1,2,3,...$ or to perform $q > 1$ independent parallel runs of data augmentation with $d = 1,2,3,....$

In the application considered here with missing values only occurring in the variable Y, and X and W referring to fully observed covariates the vector ς describes the parameters of the regression model $Y \mid X, W$. The vector $\tilde{Y}$ consists of the observed part $\tilde{Y}_{obs}$ and the missing part $\tilde{Y}_{mis}$. The indicator matrix $\tilde{I}$ reduces to a single vector indicating response on the variable Y. Under the assumption of MAR, given a current estimate $\varsigma^{(d)}$ the I-step of data augmentation reduces to

$$\tilde{Y}_{mis}^{(d+1)} \sim f(\tilde{Y}_{mis} \mid \tilde{Y}_{obs}, \tilde{X}, \tilde{W}, \varsigma^{(d)}), \tag{6.54}$$

given the current estimate of the parameters $\varsigma^{(d)}$. The P-step reduces to

$$\varsigma^{(d+1)} \sim f(\varsigma \mid (\tilde{Y}_{obs}, \tilde{Y}_{mis}^{(d+1)})', \tilde{X}, \tilde{W}) \ . \tag{6.55}$$

The implementation of data augmentation, which in the I-step imputes the missing values given a vector of parameters and in the P-step generates estimates of these unknown parameters based on the augmented data, can be viewed as a form of parameter estimation for the regression model of interest, here $f(\tilde{Y} \mid \tilde{X}, \tilde{W})$. Applying the iterative procedure until convergence the method allows drawing estimates of the parameters from their posterior distribution. The parameter estimates are used to obtain predicted values of Y and to impute the missing values for Y either by parametric random regression imputation or using forms of predictive mean matching imputation.

In particular, data augmentation, as well as other related Markov chain Monte Carlo algorithms, can be used to generate proper multiple imputation in the sense of Rubin (1987), since multiple imputations are draws from $f(\tilde{H}_{mis} \mid \tilde{H}_{obs})$, which is the limiting distribution of the I-step of data augmentation. A disadvantage of such a method can be that convergence of the algorithm might be difficult to monitor. More discussion of convergence can be found in Rosenthal (1993), Gilks, Richardson and Spiegelhalter (1996), Schafer (1997) and Tanner (1997). There is a growing literature discussing the use of data augmentation under the assumption of MAR in particular with reference to the generation of multiple imputation, see for example Rubin (1987), Schafer (1997), Tanner and Wong (1982), Heitjan and Little (1991) and Raghunathan et al. (2001). It should be noted that data augmentation has similarities to the EM-algorithm, in particular the E- and I-step and the M- and P-step are related. Both methods aim to solve a difficult missing-data problem by repeatedly applying tractable complete-data methods. In comparison to the EM-algorithm that is

deterministic and converges to a point in the parameter space, a Markov Chain Monte Carlo algorithm is stochastic and converges to a probability distribution. Whereas the *E*-step calculates the expected values of the complete-data sufficient statistics, the *I*-step performs a random draw from the complete-data sufficient statistics. The *M*-step of the *EM*-algorithm maximises a complete-data likelihood, whereas the *P*-step simulates a random draw from a complete-data posterior. *EM* therefore provides only point estimates of the unknown parameters whereas data augmentation provides random draws from their joint posterior distribution.

6.3.5.3 Data Augmentation Under the CME Assumption

The aim is to derive an imputation method based on data augmentation defined under the assumption of common measurement error instead of the commonly used assumption of MAR. One approach for imputation in the context of a measurement error model using Gibbs sampling is discussed in Carroll, Ruppert and Stefanski (1995). Their method assumes that a variable of interest Y cannot be observed but instead the surrogate X, measuring Y with measurement error, is fully observed and has an independent replicate T at every observation. The way Gibbs sampling is used in this context is to treat the unobserved values for Y as unobserved random effects and therefore as unknown parameters. The likelihood is factorised into three factors and multiplied with the prior to obtain the joint distribution of the data and the unknown parameters. Gibbs sampling is then applied to all the unknown parameters, i.e. the unknown coefficients of the underlying regression models, and the unknown values for Y. The method is illustrated for two examples. A disadvantage of the method is the computational burden that arises from drawing from the posterior distribution of the unobserved Y's, particularly if the sample size is large. In addition, an independent replicate T is assumed. Glynn, Laird and Rubin (1993) suggest the use of multiple imputation under nonignorable nonresponse in the presence of follow-up data such that information on the nonrespondents is available on a randomised subset of the sample. However, these methods are not adapted here.

The implementation of data augmentation, which in the *I*-step imputes the missing values given a vector of parameters and in the *P*-step generates estimates of these unknown parameters based on the augmented data, can be viewed, as mentioned in section 6.3.5.2, as a form of parameter

estimation for the regression model of interest. An approach that focuses on parameter estimation using data augmentation in the presence of variables measured with error is described by Kuha (1997). He considers the case of estimating parameters of regression models where one or more of the covariates are measured with error and the true values are regarded as missing. The data augmentation procedure imputes the missing values in the *I*-step and therefore creates a complete dataset with variables measured without error. These models are fitted in the posterior step and estimates of the coefficients referring to the correctly measured covariates are obtained. However, in this approach it is assumed that data from a validation study is available to obtain information about the measurement error. Other methods have been proposed for estimating the parameters of regression models where the nonresponse depends on the variable subject to missing data. One method under such a nonignorable nonresponse mechanism has been proposed by Greenlees, Reece and Zieschang (1982). An application to income data is discussed. The method estimates the parameters of a linear regression model predicting Y and the parameters of a nonresponse model dependent on Y analytically using maximum likelihood estimation. A model for Y in the population is assumed, regressed on a vector of auxiliary variables μ of the form $y_i = \mu_i\beta + \varepsilon_i$. In addition, the model of response $P(I_i = 1 | y_i, \mathbf{z}_i)$ is specified, where $\mathbf{z}_i$ is a vector of auxiliary variables. Greenlees, Reece and Zieschang (1982) specify the likelihood function for the sample where one product of the likelihood refers to the respondents and one to the nonrespondents. The likelihood function incorporates both models of interest. Therefore maximum likelihood estimates of the parameters are obtained by numerically maximizing the logarithm of this function given Y for the nonrespondents and μ and Z for the whole sample. Based on the predictions of the two models under the estimated parameters imputed values are obtained using an acceptance-rejection algorithm based on the model of the nonresponse. However, the approach by Greenlees, Reece and Zieschang (1982) has not been used here. A comparison of this approach with the method proposed in this section using data augmentation is a possible area of further research and will be discussed in chapter 7.

In the following the data augmentation based on the assumption of common measurement error reflecting a nonignorable nonresponse mechanism is derived. The difference to the simpler case under MAR is that the nonresponse model for the variable *I*, the indicator of nonresponse, as well as the parameter for the nonresponse mechanism ψ need to be taken into account. The algorithm

of data augmentation under nonignorable nonresponse may be motivated by the following two expressions, similarly to the case of data augmentation under the MAR assumption. The posterior density of the missing data can be written as

$$f(\tilde{H}_{mis} \mid \tilde{H}_{obs}, \tilde{I}) = \int\int f(\tilde{H}_{mis}, \varsigma, \psi \mid \tilde{H}_{obs}, \tilde{I}) d\varsigma d\psi ,$$

$$= \int\int f(\tilde{H}_{mis} \mid \tilde{H}_{obs}, \tilde{I}, \varsigma, \psi) f(\varsigma, \psi \mid \tilde{H}_{obs}, \tilde{I}) d\varsigma d\psi . \qquad (6.56)$$

On the other hand the desired observed posterior distribution can be written as

$$f(\varsigma, \psi \mid \tilde{H}_{obs}, \tilde{I}) = \int f(\varsigma, \psi \mid \tilde{H}_{mis}, \tilde{H}_{obs}, \tilde{I}) f(\tilde{H}_{mis} \mid \tilde{H}_{obs}, \tilde{I}) d\tilde{H}_{mis} . \qquad (6.57)$$

The integral in (6.56) contains the density $f(\varsigma, \psi \mid \tilde{H}_{obs}, \tilde{I})$ and in return the integral in (6.57) contains the expression $f(\tilde{H}_{mis} \mid \tilde{H}_{obs}, \tilde{I})$. The expression (6.56) motivates the *I*-step and (6.57) motivates the *P*-step of data augmentation incorporating the nonresponse mechanism. Note that $f(\tilde{H}_{mis} \mid \tilde{H}_{obs}, \tilde{I})$ depends on the observed response pattern $\tilde{I}$.

In the following, notation such as $f(\tilde{Y} \mid \tilde{X}, \tilde{W}, \zeta)$ is used to denote the distribution $\tilde{Y}$ given $\tilde{X}$ and $\tilde{W}$ with ζ a column-vector of parameters. The covariates W are treated as fixed constants (see also Carroll, Ruppert and Stefanski, 1995, p. 167). Thus, the likelihood of interest is the conditional density of $\tilde{Y}$, $\tilde{X}$ and $\tilde{I}$ given $\tilde{W}$ and the parameters ζ, i.e. $f(\tilde{Y}, \tilde{X}, \tilde{I} \mid \tilde{W}, \zeta)$. The posterior is the conditional density of the parameters given all the data, $f(\zeta \mid \tilde{Y}, \tilde{X}, \tilde{I}, \tilde{W})$. The joint density of the data and the parameters is given as $f(\tilde{Y}, \tilde{X}, \tilde{I}, \zeta \mid \tilde{W})$, which can be factorised into two factors referring to the likelihood and the prior. Note that for simplicity the vector of parameters ζ is not further specified here. In the derivations for the *P*-step we will specify ζ in more detail. Using data augmentation an imputation step and a posterior step need to be evaluated. Before describing the *I*- and the *P*-step under this alternative assumption in more detail two possible factorisations of the likelihood are discussed, which will form the basis of the specifications of the *I*- and the *P*-step in the data augmentation procedure. The following theory is derived using Y and X for simplicity. In the actual application to LFS data and in the simulation study functions of the variables, such as $\ln(Y)$ and $\ln(X)$, are used.

a.) Factorisation of the Likelihood Function

First, possible factorisations of the likelihood are investigated. In the *I*-step the aim is to impute the missing values of $\tilde{Y}$ by drawing imputed values from the predictive distribution $f(\tilde{Y} \mid \tilde{X}, \tilde{I} = 0, \tilde{W}, \zeta)$, where

$$f(\tilde{Y} \mid \tilde{X}, \tilde{I}, \tilde{W}, \zeta) = \frac{f(\tilde{Y}, \tilde{X}, \tilde{I} \mid \tilde{W}, \zeta)}{f(\tilde{X}, \tilde{I} \mid \tilde{W}, \zeta)}, \tag{6.58}$$

and the numerator is the likelihood function of interest $f(\tilde{Y}, \tilde{X}, \tilde{I} \mid \tilde{W}, \zeta)$. In the *P*-step the posterior distribution is expressed as the product of the likelihood and the prior. Let $f(\zeta)$ denote the prior distribution for the vector of parameters ζ. We have

$$f(\zeta \mid \tilde{Y}, \tilde{X}, \tilde{I}, \tilde{W}) \propto f(\tilde{Y}, \tilde{X}, \tilde{I} \mid \tilde{W}, \zeta) f(\zeta), \tag{6.59}$$

assuming $f(\zeta \mid W) = f(\zeta)$. Since the likelihood $f(\tilde{Y}, \tilde{X}, \tilde{I} \mid \tilde{W}, \zeta)$ is an important component in both the imputation and the posterior step, possible factorisations of the likelihood are investigated, which will form the basis of the *I*- and the *P*-step. The aim is to simplify the expression using the assumption of common measurement error. One possibility is the following.

$$\begin{aligned} f(\tilde{Y}, \tilde{X}, \tilde{I} \mid \tilde{W}, \zeta) &= f(\tilde{Y} \mid \tilde{W}, \zeta) f(\tilde{I}, \tilde{X} \mid \tilde{Y}, \tilde{W}, \zeta) \\ &= f(\tilde{Y} \mid \tilde{W}, \zeta) f(\tilde{X} \mid \tilde{Y}, \tilde{I}, \tilde{W}, \zeta) f(\tilde{I} \mid \tilde{Y}, \tilde{W}, \zeta) \\ &= f(\tilde{Y} \mid \tilde{W}, \zeta) f(\tilde{X} \mid \tilde{Y}, \tilde{W}, \zeta) f(\tilde{I} \mid \tilde{Y}, \tilde{W}, \zeta), \end{aligned} \tag{6.60}$$

where the first factor $f(\tilde{Y} \mid \tilde{W}, \zeta)$ represents a model for the variable Y subject to missing data, the second factor $f(\tilde{X} \mid \tilde{Y}, \tilde{W}, \zeta)$ the measurement error model and the last factor $f(\tilde{I} \mid \tilde{Y}, \tilde{W}, \zeta)$ a model for the nonresponse being dependent on $\tilde{Y}$ and $\tilde{W}$. This factorisation might be attractive because of the incorporation of the assumed measurement error model $f(\tilde{X} \mid \tilde{Y}, \tilde{W}, \zeta)$. Alternatively, the decomposition of the likelihood using the assumption of CME can be carried out as follows.

$$f(\tilde{Y}, \tilde{X}, \tilde{I} \mid \tilde{W}, \zeta) = f(\tilde{Y}, \tilde{X} \mid \tilde{W}, \zeta) f(\tilde{I} \mid \tilde{Y}, \tilde{X}, \tilde{W}, \zeta)$$

$$= f(\tilde{Y} \mid \tilde{X}, \tilde{W}, \zeta) f(\tilde{X} \mid \tilde{W}, \zeta) f(\tilde{I} \mid \tilde{Y}, \tilde{X}, \tilde{W}, \zeta)$$

$$= f(\tilde{Y} \mid \tilde{X}, \tilde{W}, \zeta) f(\tilde{X} \mid \tilde{W}, \zeta) f(\tilde{I} \mid \tilde{Y}, \tilde{W}, \zeta), \tag{6.61}$$

where the first factor $f(\tilde{Y} \mid \tilde{X}, \tilde{W}, \zeta)$ represents a model for the variable $\tilde{Y}$ subject to nonresponse, the second factor $f(\tilde{X} \mid \tilde{W}, \zeta)$ a model for the variable $\tilde{X}$ measured with measurement error and the factor $f(\tilde{I} \mid \tilde{Y}, \tilde{W}, \zeta)$ represents a model for the nonresponse in the same way as in (6.60). Other decompositions can be thought of. However, it is not possible to use the assumption of CME for simplification such that only the factorisations (6.60) and (6.61) are of interest here. These decompositions play a major role in the calculation of the *I*-step and the *P*-step for data augmentation under the CME-assumption. First, the implementation of the imputation step is discussed under both factorisations.

b.) The Imputation Step

In the *I*-step the aim is to draw imputed values from the predictive distribution $f(\tilde{Y}_{mis} \mid \tilde{X}, \tilde{I}, \tilde{W}, \zeta)$, the posterior distribution of the missing values of $\tilde{Y}$. Under the first decomposition (6.60) the predictive distribution can be simplified as follows for some unit j for which $I_j = 0$.

$$f(y_j \mid x_j, I_j = 0, \mathrm{w}_j, \zeta) = \frac{f(y_j, x_j, I_j = 0 \mid \mathrm{w}_j, \zeta)}{f(I_j = 0, x_j \mid \mathrm{w}_j, \zeta)}$$

$$= \frac{f(y_j \mid \mathrm{w}_j, \zeta) f(x_j \mid y_j, \mathrm{w}_j, \zeta) f(I_j = 0 \mid y_j, \mathrm{w}_j, \zeta)}{f(I_j = 0, x_j \mid \mathrm{w}_j, \zeta)}$$

$$\propto f(y_j \mid \mathrm{w}_j, \zeta)[f(x_j \mid y_j, \mathrm{w}_j, \zeta) f(I_j = 0 \mid y_j, \mathrm{w}_j, \zeta)]. \tag{6.62}$$

This factorisation can be interpreted for the *I*-step as drawing possible values from the distribution $f(y_j \mid \mathrm{w}_j, \zeta)$ and performing rejection sampling based on the additional term $f(x_j \mid y_j, \mathrm{w}_j, \zeta) f(I_j = 0 \mid y_j, \mathrm{w}_j, \zeta)$. A similar procedure using an acceptance-rejection algorithm within the *I*-step of data augmentation is proposed by Heitjan and Rubin (1990).

In (6.62) the model $f(y_j \mid \mathrm{w}_j, \zeta)$, however, does not make use of the derived variable X when predicting Y, which is regarded as desirable and was generally found to improve the fit of the model for several choices of W. In addition, performing rejection sampling based on $f(x_j \mid y_j, \mathrm{w}_j, \zeta) f(I_j = 0 \mid y_j, \mathrm{w}_j, \zeta)$ requires the use of a transformation function to obtain values between zero and one. A simplified procedure, however, can be achieved by using the second factorisation (6.61). With this decomposition the predictive distribution can be simplified as follows.

$$
\begin{aligned}
f(y_j \mid x_j, I_j = 0, \mathrm{w}_j, \zeta) &= \frac{f(y_j, x_j, I_j = 0 \mid \mathrm{w}_j, \zeta)}{f(I_j = 0, x_j \mid \mathrm{w}_j, \zeta)} \\
&= f(y_j \mid x_j, \mathrm{w}_j, \zeta) \frac{f(x_j \mid \mathrm{w}_j, \zeta) f(I_j = 0 \mid y_j, \mathrm{w}_j, \zeta)}{f(I_j = 0, x_j \mid \mathrm{w}_j, \zeta)} \\
&= f(y_j \mid x_j, \mathrm{w}_j, \zeta) \frac{f(x_j \mid \mathrm{w}_j, \zeta) f(I_j = 0 \mid y_j, \mathrm{w}_j, \zeta)}{f(x_j \mid \mathrm{w}_j, \zeta) f(I_j = 0 \mid x_j, \mathrm{w}_j, \zeta)} \\
&= f(y_j \mid x_j, \mathrm{w}_j, \zeta) \frac{f(I_j = 0 \mid y_j, \mathrm{w}_j, \zeta)}{f(I_j = 0 \mid x_j, \mathrm{w}_j, \zeta)} \\
&\propto f(y_j \mid x_j, \mathrm{w}_j, \zeta) f(I_j = 0 \mid y_j, \mathrm{w}_j, \zeta), \qquad (6.63)
\end{aligned}
$$

where the model $f(I_j = 0 \mid x_j, \mathrm{w}_j, \zeta)$ has been omitted since it does not involve the unknown values of Y. For the imputation step the decomposition (6.63) can be interpreted as follows. Given a current vector of parameters $\zeta^{(d)}$ a possible imputed value for nonrespondent j, denoted $\hat{y}_j^{(d+1)^*}$, is drawn from the distribution $f(y_j \mid x_j, \mathrm{w}_j, \zeta)$, $\hat{y}_j^{(d+1)^*} \sim f(y_j \mid x_j, \mathrm{w}_j, \zeta^{(d)})$. Rejection sampling is performed based on the distribution $f(I_j = 0 \mid y_j, \mathrm{w}_j, \zeta)$, which means that the value $\hat{y}_j^{(d+1)^*}$ is accepted for imputation with probability $f(I_j = 0 \mid \hat{y}_j^{(d+1)^*}, \mathrm{w}_j, \zeta^{(d)}) = \hat{\varrho}_j^{(d+1)^*}$, where $\hat{\varrho}_j^{(d+1)^*}$ denotes the probability of nonresponse. If accepted we set $\hat{y}_j^{(d+1)^*} = \hat{y}_j^{(d+1)}$, where $\hat{y}_i^{(d+1)}$ is the imputed value for nonrespondent j in iteration $d+1$. Note that in each *I*-step only one value is imputed for each nonrespondent. The *I*-step proposed in (6.63) has therefore a simple interpretation and is computationally easier to implement than (6.62). It has the advantage that the model for the variable X, $f(x_j \mid \mathrm{w}_j, \zeta)$ in (6.61), does not need to be fitted, which means that

the method is robust against misspecification of the assumptions made about this distribution. The model for the variable Y, $f(y_j \mid x_j, \mathrm{w}_j, \zeta)$, includes the variable X as well as other covariates W, which is thought to be of higher predictive power than the model $f(y_j \mid \mathrm{w}_j, \zeta)$. Another advantage is that the distribution $f(I_j = 0 \mid y_j, \mathrm{w}_j, \zeta)$ already generates probabilities between zero and one that can be used without further transformation for the acceptance-rejection algorithm. Because of the advantages described here it was decided to use the decomposition in (6.63) for the implementation of the *I*-step. Note that the assumption of common measurement error in (6.63) is necessary to identify the nonresponse model and to incorporate information on the measurement error (see also Kuha, 1997). If the model $f(I_j = 0 \mid y_j, x_j, \mathrm{w}_j, \zeta)$, including the derived variable X, is used in the acceptance-rejection procedure the algorithm does not converge as shown in section 6.3.5.4, since the data does not identify the nonresponse model. Therefore there is no basis for inference (see also Schafer, 1997, p. 82). Note that both (6.62) and (6.63) reflect a simple extension to the *I*-step used under the assumption of MAR where values are drawn from the predictive distribution $f(y_j \mid x_j, \mathrm{w}_j, \zeta)$ without the addition of rejection sampling.

c.) The Posterior Step

In the following the *P*-step is discussed drawing estimates of the parameters from the posterior distribution. Since the *I*-step has a simple interpretation under the second factorisation of the likelihood we concentrate on factorisation (6.61). First, the vector of parameters ζ is specified further. Under the second decomposition (6.61) the vector of parameters ζ is interpreted as $(\zeta_1', \zeta_2', \psi')'$, where ζ_1 is the column-vector of parameters of the predictive distribution of Y, $f(\tilde{Y} \mid \tilde{X}, \tilde{W}, \zeta_1)$, ζ_2 denotes the column-vector of parameters of the predictive distribution of X, $f(\tilde{X} \mid \tilde{W}, \zeta_2)$, and as introduced in chapter 1 the term ψ refers to the column-vector of parameters of the response model, here $f(\tilde{I} \mid \tilde{Y}, \tilde{W}, \psi)$. Note that the parameters ζ_1, ζ_2 and ψ refer to the complete-data models respectively. We also introduce the notation $\zeta_{1,mis}$ for the vector of parameters of the model $f(\tilde{Y}_{mis} \mid \tilde{X}, \tilde{I}, \tilde{W}, \zeta_{1,mis})$, which is based on the nonrespondents only. Let $f(\zeta)$ denote the prior distribution for the parameters ζ. The complete-data posterior distribution can then be expressed as

$$f(\zeta \mid \tilde{Y}, \tilde{X}, \tilde{I}, \tilde{W}) \propto f(\tilde{Y}, \tilde{X}, \tilde{I} \mid \tilde{W}, \zeta) f(\zeta)$$

$$= f(\tilde{Y} \mid \tilde{X}, \tilde{W}, \zeta_1) f(\tilde{X} \mid \tilde{W}, \zeta_2) f(\tilde{I} \mid \tilde{Y}, \tilde{W}, \psi) f(\zeta_1, \zeta_2, \psi)$$

$$= f(\tilde{Y} \mid \tilde{X}, \tilde{W}, \zeta_1) f(\zeta_1)\, f(\tilde{X} \mid \tilde{W}, \zeta_2) f(\zeta_2)\, f(\tilde{I} \mid \tilde{Y}, \tilde{W}, \psi) f(\psi), \tag{6.64}$$

assuming that the parameters ζ_1, ζ_2 and ψ are distinct (see definition 1.3), such that the prior distribution $f(\zeta)$ factors as,

$$f(\zeta) = f(\zeta_1, \zeta_2, \psi) = f(\zeta_1) f(\zeta_2) f(\psi). \tag{6.65}$$

The rationale for the assumption that the parameters of the data model, $\zeta^* = (\zeta_1', \zeta_2')'$, and the nonresponse model, ψ, are distinct is that knowing the parameters of a model for the data will provide little information about the parameters of the nonresponse model and vice versa. The assumption of independent parameters of the data and the nonresponse is often made in the literature and will be used here (Heitjan and Rubin, 1990, p. 308; Schafer, 1997, p. 18; Carroll, Ruppert and Stefanski, 1997, p. 173). In terms of the prior for $\zeta^* = (\zeta_1', \zeta_2')'$ it will be shown in equation (6.75) that the prior falls into independent priors for ζ_1 and ζ_2. Note that in the example discussed here the likelihood function has a simple interpretation as the product of three complete-data likelihood functions and each parameter only appears in one of the three parts of the likelihood, as shown in (6.69). The posterior therefore factors into a series of independent posteriors $f(\zeta_1 \mid \tilde{Y}, \tilde{X}, \tilde{W})$, $f(\zeta_2 \mid \tilde{X}, \tilde{W})$ and $f(\psi \mid \tilde{Y}, \tilde{I}, \tilde{W})$. Examples where the likelihood can be expressed as a product of several complete-data likelihoods, where the parameters of the two factors are not distinct, are given in Schafer (1997).

To proceed we need to compute (6.64) making assumptions about the underlying distributions of the data and choosing adequate priors. Before discussing the choice of priors, a specification of the likelihood function is given. The assumption of normality is commonly made in the literature for continuous data (Heitjan and Rubin, 1990; Schafer, 1997). In particular, classical linear regression analysis assumes conditional normality of the response variable given linear functions of the predictors, which is the conditional distribution implied by a multivariate normal model for all the variables. We assume for some unit i that

$$f(y_i \mid x_i, \mathrm{w}_i, \zeta_1) \sim N(\eta_i \beta; \sigma^2_{Y|X,W}), \tag{6.66}$$

where η_i is a vector of covariates, functions of the derived variable x_i and other covariates w_i, β is a vector of coefficients and $\sigma^2_{Y|X,W}$ denotes the conditional variance of Y given X and W. The vector of parameters is $\zeta_1 = (\beta', \sigma^2_{Y|X,W})'$. We further assume that

$$f(x_i \mid \mathrm{w}_i, \zeta_2) \sim N(\mu_i \alpha; \sigma^2_{X|W}), \tag{6.67}$$

where μ_i is a vector of functions of the covariates w_i, α is a vector of coefficients and $\sigma^2_{X|W}$ denotes the conditional variance of X given W. We have $\zeta_2 = (\alpha', \sigma^2_{X|W})'$. The above assumptions for regression models predicting a continuous dependent variable from a vector of continuous and discrete independent variables in the context of data augmentation have been made by Raghunathan et al. (2001) and by Heitjan and Rubin (1990) for heaped, rounded and truncated data. Greenlees, Reece and Zieschang (1982) use these assumptions for estimating regression models for income data in the presence of nonignorable nonresponse. The validity of these assumptions in the LFS example based on the variables $\ln(Y)$ and $\ln(X)$ rather than Y and X is discussed in the following before proceeding with the calculations of the likelihood.

For the validity of the imputation method it is important to assess to what extent the assumptions made in (6.66) and (6.67) for models $f(y_i \mid x_i, \mathrm{w}_i, \zeta_1)$ and $f(x_i \mid \mathrm{w}_i, \zeta_2)$ might hold in reality. Since the assumptions refer to respondents and nonrespondents, the validity of these distributional assumptions is strictly speaking untestable. Note that the imputation method does not rely on the correct specification of the model $f(x_i \mid \mathrm{w}_i, \zeta_2)$ since under factorisation (6.63) this part of the model cancels out. We therefore concentrate on the validity of the assumptions for the distribution $f(y_i \mid x_i, \mathrm{w}_i, \zeta_1)$.

In the LFS application we use $\ln(Y)$ and $\ln(X)$ and refer to the model $f(\ln(y_i) \mid \ln(x_i), \mathrm{w}_i, \zeta_1)$. The regression model for $\ln(Y)$, specified in table 2.1, is applied to the LFS quarter March-May 2000. The analysis is, however, only based on respondent data. For regression coefficients and model diagnostics see appendix A2.1. The histogram of the residuals given in figure A2.1.2 and the normal probability plot in figure A2.1.3 in the appendix show a distribution which approximates the normal distribution reasonably well. We therefore conclude that the assumption of normality, using the *ln* transformation on both variables Y and X, seems adequate. To evaluate the assumption of constant variance of the residuals the plot of the residuals versus fitted values given

in figure A2.1.1 is investigated. The plot shows a random spread with no obvious pattern. However, some outlying values seem to indicate an increase in the variance with increasing predicted values for Y. Table 6.18 gives the standard deviation within classes of the predicted values. We observe an increase in the standard deviation with increasing predicted values of $\ln(Y)$. However, note that some classes are based on only a small number of respondents particularly at the bottom and top end of the distribution.

Class	Standard Deviation
1	0.21
2	0.15
3	0.12
4	0.14
5	0.14
6	0.16
7	0.18
8	0.19
9	0.21
10	0.22
11	0.23
12	0.24
13	0.26

Table 6.18: Standard deviation of the residuals of the imputation model.

To address the problem of a possible departure from the assumption of non-constant variance the use of hot deck imputation based on the predicted values instead of parametric random regression imputation is considered in the *I*-step of data augmentation. This approach is evaluated in sections 6.3.4.4 and 6.3.4.5 and it is found that the method seems to be robust to departures from the assumption of constant variance.

Schafer (1997) emphasizes that real data often deviates from the assumption of normality and that the underlying probability model is only an approximation. He stresses the fact that in many cases even when the underlying data are not normal the assumption of normality is still useful, particularly since transformations can be applied to the data to make the assumption more

realistic. In addition, since the aim of the imputation procedure and the models specified is prediction the family of linear models is recommended (Draper and Smith, 1981, ch. 14; Heitjan and Rubin, 1990). Since often the data does not conform to normality it is important to evaluate whether the imputation procedure using data augmentation is robust to departures from the underlying modelling assumptions. Schafer (1997, section 6.4) demonstrates in a simulation experiment the robustness of multiple imputation to moderate departures of the normal model. He emphasizes that even if the observed data are nonnormal it is sensible to use a normal model to generate multiple imputations. Rubin and Schenker (1986) and Rubin (1987) provide similar evidence.

We now proceed with the calculations for the likelihood. For the response model, also called missing data mechanism, the following notation is introduced. Note that the distribution required for the imputation step in the data augmentation procedure refers to the probability of nonresponse rather than response. Let

$$f(I_i = 1 \mid y_i, \mathrm{w}_i, \psi) = G(\tau_i \psi) = p_i, \tag{6.68}$$

where τ_i is a vector including functions of the direct variable Y and other covariates W, ψ is a vector of coefficients of the regression model modelling response, p_i denotes the probability of response and G is a function such as the logistic regression model, $G(\tau_i \psi) = \frac{\exp(\tau_i \psi)}{1+\exp(\tau_i \psi)}$.

It follows for the complete-data likelihood under the decomposition in (6.61), assuming independence across units given the parameters,

$$f(\tilde{Y}, \tilde{X}, \tilde{I} \mid \tilde{W}, \zeta) = f(\tilde{Y} \mid \tilde{X}, \tilde{W}, \zeta) f(\tilde{X} \mid \tilde{W}, \zeta) f(\tilde{I} \mid \tilde{Y}, \tilde{W}, \zeta)$$

$$= \prod_{i=1}^{n} f(y_i \mid x_i, \mathrm{w}_i, \zeta_1) f(x_i \mid \mathrm{w}_i, \zeta_2) f(I_i \mid y_i, x_i, \psi)$$

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

$$* \prod_{i=1}^{n} (\sigma^2_{X|W})^{-1/2} \exp\left\{ \frac{-1}{2\sigma^2_{X|W}} (x_i - \mu_i \alpha)^2 \right\}$$

$$*\prod_{i=1}^{n} G^{I_i}(\tau_i \psi)\{1-G(\tau_i \psi)\}^{(1-I_i)}, \tag{6.69}$$

where $\eta_i\beta = E(y_i \mid x_i, \mathrm{w}_i)$ and $\mu_i \alpha = E(x_i \mid \mathrm{w}_i)$.

We now turn to the specification of the prior distributions. In practice, priors are often chosen, at least to some extent, for reasons of computational convenience. A simple way to conduct Bayesian inference in the complete data case is to apply a class of conjugate priors such that any prior $f(\zeta)$ in the class leads to a posterior distribution that belongs to the same class as the likelihood function. A commonly used conjugate class for the multivariate normal data model is the normal inverted-Wishart family. Assuming a prior distribution based on the inverted Wishart distribution and computing the complete-data likelihood function leads to a complete-data posterior, which is normal inverted Wishart (Schafer, 1997, 5.2.2). In the case of no strong prior information about ζ a noninformative prior is used. Also under a noninformative prior the complete-data posterior is normal inverted Wishart, where the hyperparameters of the resulting posterior distribution depend on the chosen prior. More on the choice of priors can be found in Box and Tiao (1992), Bedrick, Christensen and Johnson (1996) and Gelman et al. (1998). In the special case of only one variable being subject to missing data the inverted Wishart reduces to a scaled inverted chisquare distribution. Note that in general this family is not conjugate when some of the data is missing. Only in special cases the observed-data posterior is tractable under a normal inverted Wishart prior. However, since data augmentation only relies on the tractability of the complete-data posterior the method is easy to implement under such a family of priors. In the special case that the data has a monotone missing-data pattern and if the parameters are distinct and if the likelihood factors into different components according to the parameters of interest, the posterior of interest can be expressed as a combination of a normal and a scaled inverted chisquare distribution. This case will be discussed here. (Note: A missing-pattern is said to be *monotone* if, whenever an element h_{ij} of the data matrix $\tilde{H}$ is missing, h_{ik} is also missing for all $k > j$ (Little and Rubin, 2002).)

To specify the prior $f(\zeta)$, in the example discussed here, we make assumptions about $f(\psi)$ and $f(\zeta_1,\zeta_2)$, i.e. a prior for the vectors of parameters ζ_1 and ζ_2 is specified jointly. It will be shown that $f(\zeta_1,\zeta_2)$ falls into separate priors for ζ_1 and ζ_2. In the following, noninformative priors are

used. For large samples and not too many parameters, as it is the case in this application, the noninformative prior seems useful, whereas for small sample sizes the prior distribution is more important (Gelman et al., 1998). First, the prior for ζ_1 and ζ_2 and the resulting posterior distribution for the two vectors of parameters are described. Let υ denote the parameters of the joint distribution of $\tilde{X}$ and $\tilde{Y}$ given $\tilde{W}$, a bivariate normal distribution, where υ includes the variance-covariance matrix $\sum = \begin{pmatrix} \sigma^2_{X|W} & \sigma_{X,Y|W} \\ \sigma_{Y,X|W} & \sigma^2_{Y|W} \end{pmatrix}$. We have

$$f(\tilde{X},\tilde{Y}\,|\,W,\upsilon) = f(\tilde{Y}\,|\,\tilde{X},\tilde{W},\zeta_1) f(\tilde{X}\,|\,\tilde{W},\zeta_2), \tag{6.70}$$

where υ and $\zeta^* = (\zeta_1',\zeta_2')'$ are functions of each other (for further information on these functions see Schafer, 1997, section 5.2.4). Following the derivations in Schafer (1997) and Box and Tiao (1992) a prior distribution for ζ^* can be specified as follows. A commonly used noninformative prior for υ in the case of complete data is

$$f(\upsilon) \propto |\sum|^{-3/2}, \tag{6.71}$$

which is the limiting form of the normal inverted Wishart density and arises by applying Jeffreys invariance principle to obtain an improper prior. One important justification for this prior with complete data is that Bayesian and frequentist inferences about the mean coincide since the highest posterior density region for the mean under this prior is identical to the classical confidence region for the mean (Schafer, 1997, p. 155). Since we are interested in the prior for ζ^* and because $\zeta^* = \zeta^*(\upsilon)$ is a one-to-one transformation the prior for ζ^* can be expressed as

$$f(\zeta^*) = f(\upsilon)\|J\|^{-1} \propto |\sum|^{-3/2} \|J\|^{-1}, \tag{6.72}$$

where $\|J\|^{-1}$ denotes the absolute value of the determinant of the Jacobian matrix for the transformation from υ to ζ^*. Following arguments in Mardia, Kent and Bibby (1979), we obtain $\|J\|^{-1} = \sigma^2_{X|W}$. A general result for determinants of matrices where $\sum_{11}$ and $\sum_{22}$ are square submatrices, says that

$$|\sum| = \begin{vmatrix} \sum_{11} & \sum_{12} \\ \sum_{21} & \sum_{22} \end{vmatrix} = \left|\sum_{11}\right| \left|\sum_{22} - \sum_{21}\sum_{11}^{-1}\sum_{12}\right|, \tag{6.73}$$

where the term $\Sigma_{22}-\Sigma_{21}\Sigma_{11}^{-1}\Sigma_{12}$ is the residual covariance matrix from the regression of the variables corresponding to Σ_{22} on the variables corresponding to Σ_{11}. Using this result we obtain the determinant of Σ, where $\Sigma=\begin{pmatrix}\sigma^2_{X|W} & \sigma_{X,Y|W}\\ \sigma_{Y,X|W} & \sigma^2_{Y|W}\end{pmatrix}$,

$$|\Sigma|=\sigma^2_{X|W}(\sigma^2_{Y|W}-\sigma_{Y,X|W}\sigma^{-2}_{X|W}\sigma_{X,Y|W})=\sigma^2_{X|W}\,\sigma^2_{Y|X,W} \tag{6.74}$$

and it follows for the prior of ζ^*

$$f(\zeta^*)\propto(\sigma^2_{X|W}\sigma^2_{Y|X,W})^{-3/2}\sigma^2_{X|W}=(\sigma^2_{X|W})^{-1/2}(\sigma^2_{Y|X,W})^{-3/2}. \tag{6.75}$$

We see that under the prior for υ as specified in (6.71) the prior distribution for ζ^* factors into independent priors for ζ_1 and ζ_2.

Since in the example discussed here, where only the variable Y is subject to missing data, the dataset has a monotone missing data structure. In the case of monotone missingness and if the parameters are distinct, which is the case here, and if the likelihood factors into different components according to the parameters of interest, as shown in (6.69), the following special case applies for calculating the required posterior distribution of the parameters. Combining prior and complete-data likelihood expressions for the complete-data posteriors $f(\zeta_1\,|\,\tilde{Y},\tilde{X},\tilde{W})$ and $f(\zeta_2\,|\,\tilde{X},\tilde{W})$ are obtained that enable us to draw estimates of the desired parameters based on complete data. The posterior distribution for $\zeta_1=(\beta',\sigma^2_{Y|X,W})'$, discarding proportionality constants, is

$$f(\zeta_1\,|\,\tilde{Y},\tilde{X},\tilde{W})\propto f(\tilde{Y}\,|\,\tilde{X},\tilde{W},\zeta_1)\,f(\zeta_1)$$

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

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

$$= (\sigma^2_{Y|X,W})^{-(\frac{n+3}{2})} \exp\left\{\frac{-1}{2\sigma^2_{Y|X,W}} \sum_{i=1}^{n} (y_i - \eta_i\beta)^2\right\}, \tag{6.76}$$

which may be expressed as in the following proposition (see also Schafer, 1997).

Proposition 6.1

The posterior distribution $f(\zeta_1 \mid \tilde{Y}, \tilde{X}, \tilde{W})$ with $\zeta_1 = (\beta', \sigma^2_{Y|X,W})'$, is the product of a multivariate normal distribution, $N(\hat{\beta}, \sigma^2_{Y|X,W}(\tilde{\eta}'\tilde{\eta})^{-1})$, and a scaled inverted chisquare distribution, $\hat{\tilde{\varepsilon}}'_Y \hat{\tilde{\varepsilon}}_Y \chi^{-2}_{n-1}$, with n-1 degrees of freedom and scaling factor $\hat{\tilde{\varepsilon}}'_Y \hat{\tilde{\varepsilon}}_Y$, where β and η are defined in (6.66) and $\tilde{\eta}$ defines the corresponding matrix, $\hat{\beta}$ is the maximum likelihood (ML) estimate based on the augmented data taking into account respondents and nonrespondents, $\hat{\beta} = (\tilde{\eta}'\tilde{\eta})^{-1}\tilde{\eta}'\tilde{Y}$ and $\hat{\tilde{\varepsilon}}_Y = \tilde{Y} - \tilde{\eta}\hat{\beta}$ based on augmented data $\tilde{Y}$.

We can therefore draw the required parameters as follows

$$\sigma^2_{Y|X,W} \mid \tilde{Y}, \tilde{W} \sim \hat{\tilde{\varepsilon}}'_Y \hat{\tilde{\varepsilon}}_Y \chi^{-2}_{n-1} \tag{6.77}$$

$$\beta \mid \sigma^2_{Y|X,W}, \tilde{Y}, \tilde{W} \sim N(\hat{\beta}, \sigma^2_{Y|X,W}(\tilde{\eta}'\tilde{\eta})^{-1}), \tag{6.78}$$

The drawing of the parameters is therefore carried out sequentially drawing first from (6.77) and then from (6.78).

Proof (proposition 6.1)

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

$$= (\sigma^2_{Y|X,W})^{-(\frac{n+3}{2})} \exp\left\{\frac{-1}{2\sigma^2_{Y|X,W}} (\tilde{Y} - \tilde{\eta}\beta)'(\tilde{Y} - \tilde{\eta}\beta)\right\}$$

$$= (\sigma^2_{Y|X,W})^{-(\frac{n+3}{2})} \exp\left\{\frac{-1}{2\sigma^2_{Y|X,W}} \left(\hat{\tilde{\varepsilon}}'_Y \hat{\tilde{\varepsilon}}_Y + (\beta - \hat{\beta})'\tilde{\eta}'\tilde{\eta}(\beta - \hat{\beta})\right)\right\}$$

$$\text{since } (\tilde{Y}-\tilde{\eta}\beta)'(\tilde{Y}-\tilde{\eta}\beta)=\hat{\tilde{\varepsilon}}_Y'\hat{\tilde{\varepsilon}}_Y+(\beta-\hat{\beta})'\tilde{\eta}'\tilde{\eta}(\beta-\hat{\beta})$$

$$=(\sigma^2_{Y|X,W})^{-1}\exp\left\{\frac{-1}{2\sigma^2_{Y|X,W}}(\beta-\hat{\beta})'\tilde{\eta}'\tilde{\eta}(\beta-\hat{\beta})\right\}(\sigma^2_{Y|X,W})^{-(\frac{n+1}{2})}\exp\left\{\frac{-1}{2\sigma^2_{Y|X,W}}\hat{\tilde{\varepsilon}}_Y'\hat{\tilde{\varepsilon}}_Y\right\}$$

$$=(\sigma^2_{Y|X,W})^{-1}\exp\left\{\frac{-1}{2\sigma^2_{Y|X,W}}(\beta-\hat{\beta})'\tilde{\eta}'\tilde{\eta}(\beta-\hat{\beta})\right\}(\sigma^2_{Y|X,W})^{-(\frac{n-1}{2})-1}\exp\left\{-\frac{1}{2}\hat{\tilde{\varepsilon}}_Y'\hat{\tilde{\varepsilon}}_Y\,\sigma^2_{Y|X,W}\right\}.$$

The posterior is therefore the product of a multivariate normal distribution, $N(\hat{\beta},\sigma^2_{Y|X,W}(\tilde{\eta}'\tilde{\eta})^{-1})$, and a scaled inverted chisquare distribution, $\hat{\tilde{\varepsilon}}_Y'\hat{\tilde{\varepsilon}}_Y\chi^{-2}_{n-1}$, with n-1 degrees of freedom and scaling factor $\hat{\tilde{\varepsilon}}_Y'\hat{\tilde{\varepsilon}}_Y$.

❑

Similarly, combining the prior for ζ_2 with the likelihood $f(\tilde{X}\,|\,\tilde{W},\zeta_2)$, it follows for the complete-data posterior of the vector of parameters ζ_2, discarding proportionality constants,

$$f(\zeta_2\,|\,\tilde{X},\tilde{W})\propto f(\tilde{X}\,|\,\tilde{W},\zeta_2)f(\zeta_2)=\left(\prod_{i=1}^{n}(\sigma^2_{X|W})^{-1/2}\exp\left\{\frac{-1}{2\sigma^2_{X|W}}(x_i-\mu_i\alpha)^2\right\}\right)(\sigma^2_{X|W})^{-1/2}$$

$$=(\sigma^2_{X|W})^{-1/2}(\sigma^2_{X|W})^{-n/2}\exp\left\{\frac{-1}{2\sigma^2_{Y|X,W}}\sum_{i=1}^{n}(x_i-\mu_i\alpha)^2\right\}$$

$$=(\sigma^2_{X|W})^{-(\frac{n+1}{2})}\exp\left\{\frac{-1}{2\sigma^2_{Y|X,W}}\sum_{i=1}^{n}(x_i-\mu_i\alpha)^2\right\}$$

$$=(\sigma^2_{X|W})^{-1/2}\exp\left\{\frac{-1}{2\sigma^2_{Y|X,W}}(\alpha-\hat{\alpha})'\tilde{\mu}'\tilde{\mu}(\alpha-\hat{\alpha})\right\}(\sigma^2_{X|W})^{-(\frac{n-2}{2}+1)}\exp\left\{-\frac{1}{2}\hat{\tilde{\varepsilon}}_X'\hat{\tilde{\varepsilon}}_X\,\sigma^2_{X|W}\right\},$$

using similar arguments as in the proof of proposition 6.1, where α and μ are defined in (6.67) and $\tilde{\mu}$ denotes the corresponding matrix, $\hat{\alpha}$ is the maximum likelihood estimate, $\hat{\alpha}=(\tilde{\mu}'\tilde{\mu})^{-1}\tilde{\mu}'\tilde{X}$ and $\hat{\tilde{\varepsilon}}_X'=\tilde{X}-\tilde{\mu}\hat{\alpha}$. We see that $f(\zeta_2\,|\,\tilde{X},\tilde{W})$ is the product of a multivariate normal and a scaled inverted chisquare distribution such that

$$\sigma^2_{X|W} \mid \tilde{W} \sim \hat{\tilde{\varepsilon}}'_X \hat{\tilde{\varepsilon}}_X \chi^{-2}_{n-2} \tag{6.79}$$

$$\alpha \mid \sigma^2_{X|W}, \tilde{W} \sim N(\hat{\alpha}, \sigma^2_{X|W}(\tilde{\mu}'\tilde{\mu})^{-1}). \tag{6.80}$$

Note that the distribution $f(\zeta_2 \mid \tilde{X}, \tilde{W})$ does not need to be considered when using factorisation (6.61). Further details for calculating the posterior distribution of the parameters of regression models can be found in Raghunathan et al. (2001) and Gelman et al. (1995). Under noninformative priors it is important to check that the resulting posterior distribution is proper, i.e. that the integral is finite. For the two examples above the posterior is proper if the sample size n is greater than the number of variables of interest, here X and Y, and if the columns of the set of the covariates involved are linearly independent (Gelman et al, 1998). These conditions are fulfilled here.

We now turn to the problem of drawing parameters for the response model $f(\tilde{I} \mid \tilde{Y}, \tilde{W}, \psi)$ in the posterior step based on complete data. Several approaches for specifying priors for binomial regression problems have been proposed. Prior specifications can either focus on the regression coefficients or on eliciting binomial probabilities. One approach is to specify the regression coefficients as the parameters of interest and to state a posterior distribution for ψ that refers to the vector of regression coefficients of the binomial regression model specified in (6.68). It is common to assume a normal or diffuse prior $f(\psi)$, which is convenient in large sample situations where the posterior of ψ is approximately normal (Bedrick, Christensen and Johnson, 1997). This approach has been discussed by Leonard (1972) and Zellner and Rossi (1984). An alternative approach is to focus on the binomial probabilities for various choices of covariate values using for example beta distributions as priors for the probabilities. A discussion of Bayesian inference where the probabilities of response are regarded as the parameters of interest can be found in Bedrick, Christensen and Johnson (1996 and 1997) and Tsutakawa and Lin (1986). The approach by Tsutakawa and Lin (1986) discusses the case for binomial regression models with only one predictor variable. This method has been extended by Bedrick, Christensen and Johnson (1996 and 1997) to generalized linear models with multiple covariates. One advantage of focussing on the probabilities is that the specification of the prior are independent of the functions used for the data, such as logistic, logit etc., since the probabilities have the same interpretation irregardless of

the link function (Tsutakawa and Lin, 1986). Other examples of specifying the binomial regression model using the probabilities rather than the regression coefficients can be found in Carroll, Ruppert and Stefanski (1995) and Schafer (1997). For the example discussed here the approach of specifying priors for the vector of regression coefficients rather than the probabilities has been applied since the vector of regression coefficients is of smaller dimension than the vector of probabilities.

To be able to draw the parameters for the response model $f(\tilde{I} \mid \tilde{Y}, \tilde{W}, \psi)$ focussing on the regression coefficients as the parameters of interest a noninformative prior is specified as done in Zellner and Rossi (1984), such that $f(\psi) \propto c$, where c is a constant. Note that the notation introduced in (6.68) refers to the parameters under the response model. It follows for the complete-data posterior of ψ

$$\begin{aligned} f(\psi \mid \tilde{Y}, \tilde{I}, \tilde{W}) &\propto f(\tilde{I} \mid \tilde{Y}, \tilde{W}, \psi) f(\psi) \\ &= f(\tilde{I} \mid \tilde{Y}, \tilde{W}, \psi)\, c \\ &\propto f(\tilde{I} \mid \tilde{Y}, \tilde{W}, \psi) \\ &= \exp(\log(f(\tilde{I} \mid \tilde{Y}, \tilde{W}, \psi))). \end{aligned} \tag{6.81}$$

Let $L(\psi) \equiv \log(f(\tilde{I} \mid \tilde{Y}, \tilde{W}, \psi))$. Following the derivations in Zellner and Rossi (1984) and expanding $L(\psi)$ in a Taylor series about the modal value of (6.81), i.e. the ML estimate $\hat{\psi}$, we obtain

$$\begin{aligned} f(\psi \mid \tilde{Y}, \tilde{I}, \tilde{W}) &\propto \exp(L(\psi)) \\ &\propto \exp\{-\frac{1}{2}(\psi - \hat{\psi})' \tilde{B} (\psi - \hat{\psi}) + R\} \\ &\propto \exp\{-\frac{1}{2}(\psi - \hat{\psi})' \tilde{B} (\psi - \hat{\psi})\} \sum_{k=0}^{\infty} \frac{1}{k!} R^k, \end{aligned} \tag{6.82}$$

where R is the remainder of the Taylor series expansion,

$$\tilde{B} = -\left[\frac{\partial^2 L(\psi)}{\partial\psi\partial\psi'}\right]_{\psi=\hat{\psi}} = \tilde{\tau}'\tilde{D}\tilde{\tau}, \tag{6.83}$$

$\tilde{\tau}$ is a matrix including functions of the direct variable $\tilde{Y}$ and other covariates $\tilde{W}$ defined in (6.68) and $\tilde{D}$ is a diagonal matrix with element

$$d_i = \left[\frac{I_i}{G_i^2} + \frac{1-I_i}{(1-G_i)^2}\right] g_i^2 - \frac{(I_i - G_i)g_i'}{G_i(1-G_i)}, \tag{6.84}$$

where
$$G_i = G(\tau_i\hat{\psi}), \tag{6.85}$$

$$g_i = \left[\frac{dG(z_i)}{dz_i}\right]_{z_i=\tau_i\hat{\psi}} = g(\tau_i\hat{\psi}), \tag{6.86}$$

$$g_i' = \left[\frac{dg(z_i)}{dz_i}\right]_{z_i=\tau_i\hat{\psi}}. \tag{6.87}$$

The function G was chosen to be a logistic model of the form

$$G(\tau_i\psi) = \frac{\exp(\tau_i\psi)}{1+\exp(\tau_i\psi)}, \tag{6.88}$$

such that we obtain for (6.85) - (6.87)

$G_i = G(\tau_i\hat{\psi}) = \dfrac{\exp(\tau_i\hat{\psi})}{1+\exp(\tau_i\hat{\psi})} = \hat{p}_i$, where $\hat{p}_i$ denotes the probability of response

$$g_i = g(\tau_i\hat{\psi}) = \frac{\exp(\tau_i\hat{\psi})[1+\exp(\tau_i\hat{\psi})] - \exp(\tau_i\hat{\psi})^2}{[1+\exp(\tau_i\hat{\psi})]^2} = \frac{\exp(\tau_i\hat{\psi})}{1+\exp(\tau_i\hat{\psi})}\frac{[1+\exp(\tau_i\hat{\psi}) - \exp(\tau_i\hat{\psi})]}{1+\exp(\tau_i\hat{\psi})}$$

$$= \frac{\exp(\tau_i\hat{\psi})}{1+\exp(\tau_i\hat{\psi})}[1 - \frac{\exp(\tau_i\hat{\psi})}{1+\exp(\tau_i\hat{\psi})}] = \hat{p}_i(1-\hat{p}_i) \text{ and}$$

$g_i' = \hat{p}_i(1-\hat{p}_i)^2 + \hat{p}_i^2(1-\hat{p}_i)$,

where g and g' denote the first and second derivative of G and $\hat{p}_i = G(\tau_i\hat{\psi})$ the predicted probability of response based on the ML estimate $\hat{\psi}$. Based on the first order approximation the posterior in (6.82) can be expressed as

$$f(\psi \mid \tilde{Y}, \tilde{I}, \tilde{W}) \propto \exp\{-\frac{1}{2}(\psi - \hat{\psi})'\tilde{B}(\psi - \hat{\psi})\}, \tag{6.89}$$

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

$$\psi \sim N(\hat{\psi}, \tilde{B}^{-1}). \tag{6.90}$$

Further details of the calculations can be found in Zellner and Rossi (1984). The posterior of ψ under an informative prior have also been discussed by Zellner and Rossi (1984). If the prior is chosen to be a normal distribution it follows that the posterior is also normal and under large sample conditions the mean of the posterior is the ML estimate $\hat{\psi}$.

d.) Summary of *I*- and *P*-Step

To summarize, the *I*-step and the *P*-step of data augmentation based on the common measurement error model under the factorisation (6.61) are as follows. Note that $\zeta_{1,mis}$ denotes the parameter of the model $f(y_j \mid x_j, I_j = 0, \mathrm{w}_j, \zeta_{1,mis})$ based on the nonrespondents only, whereas the parameters ζ_1 and ψ refer to the complete-data models respectively, and in general $\zeta_{1,mis} \neq \zeta_1$. Also, ψ refers to the parameters of the response model such that $(-\psi)$, the parameters with the opposite sign, refers to the parameters of the nonresponse model $f(I_j = 0 \mid y_j, \mathrm{w}_j, (-\psi))$. Let D, $d = 1,...,D$ denote the number of iterations within the data augmentation procedure. Given a current estimate of the parameters $\zeta^{(d)}$, where $\zeta^{(d)} = (\zeta_1^{(d)'}, \psi^{(d)'})'$, we have

I-step:

$$y_j^{(d+1)} \sim f(y_j \mid x_j, I_j = 0, \mathrm{w}_j, \zeta_{1,mis}^{(d)}) \propto f(y_j \mid x_j, \mathrm{w}_j, \zeta_1^{(d)}) f(I_j = 0 \mid y_j, \mathrm{w}_j, (-\psi)^{(d)}) \tag{6.91}$$

i.e.

1.) draw $\hat{y}_j^{(d+1)^*} \sim f(y_j \mid x_j, \mathrm{w}_j, \zeta^{(d)})$ for nonrespondent $j \in \bar{r}$

2.) accept this value for imputation with probability

$$f(I_j = 0 \mid y_j^{(d+1)^*}, \mathrm{w}_j, (-\psi)^{(d)}) = 1 - f(I_j = 1 \mid y_j^{(d+1)^*}, \mathrm{w}_j, \psi^{(d)}) = \hat{\varrho}_j^{(d+1)^*}.$$

If the value $\hat{y}_j^{(d+1)^*}$ is accepted for imputation set $\hat{y}_j^{(d+1)^*} = \hat{y}_j^{(d+1)}$, where $\hat{y}_j^{(d+1)}$ is the imputed value for the nonrespondent j in iteration $d+1$. If the value is rejected return to step 1.).

P-step:

$$\zeta_1^{(d+1)} = (\beta^{(d+1)'}, \sigma^2_{Y|X,W}{}^{(d+1)})' \sim f(\zeta_1 \mid (\tilde{Y}'_{obs}, \tilde{Y}^{(d+1)'}_{mis})', \tilde{X}, \tilde{W}), \text{ and}$$

$$\psi^{(d+1)} \sim f(\psi \mid (\tilde{Y}'_{obs}, \tilde{Y}^{(d+1)'}_{mis})', \tilde{I}, \tilde{W}) \tag{6.92}$$

where $\sigma^2_{Y|X,W}{}^{(d+1)} \mid (\tilde{Y}'_{obs}, \tilde{Y}^{(d+1)'}_{mis})', \tilde{W} \sim \hat{\tilde{\varepsilon}}'_Y \hat{\tilde{\varepsilon}}_Y \chi^{-2}_{n-1}$,

$\beta^{(d+1)} \mid \sigma^2_{Y|X,W}{}^{(d+1)}, (\tilde{Y}'_{obs}, \tilde{Y}^{(d+1)'}_{mis})', \tilde{W} \sim N(\hat{\beta}, \sigma^2_{Y|X,W}{}^{(d+1)} (\tilde{\eta}'\tilde{\eta})^{-1})$

and $\psi^{(d+1)} \sim N(\hat{\psi}, \tilde{B}^{-1})$.

Since data augmentation is based on parametric random regression imputation, the use of predictive mean matching imputation is proposed in the *I*-step to relax distributional assumptions. This approach has the advantage that actually observed values are imputed that may preserve the shape of the distribution, for example preserving truncation, heaping and rounding effects. It also provides a tool to compensate for departures of constant variance of the residuals. Two forms of predictive mean matching imputation are implemented, hot deck imputation within classes and nearest neighbour imputation, where the classes and the nearest neighbours are defined based on the predictions of the regression model for Y. The acceptance-rejection procedure based on the probability $\varrho_j = 1 - f(I_j = 1 \mid y_j, \mathrm{w}_j, \psi)$ is implemented using a weighted bootstrap method as described in section 6.3.4, since classical rejection sampling requires being able to generate a large number of possible imputed values. However, under hot deck imputation within classes and nearest neighbour imputation the number of values that can be chosen for imputation is limited

due to the definition of the classes and the nearest neighbours. Under hot deck imputation within classes Q donor values, denoted $\hat{y}^{*}_{j1},...,\hat{y}^{*}_{jQ}$, are selected with simple random sampling without replacement for nonrespondent j from the same class. The value for imputation, $\hat{y}_j^{(d+1)}$, is sampled out of the Q possible values with probabilities

$$\hat{\varrho}_{jq}^{(d+1)^*} = f(I_j = 0 \mid y_{jq}^{(d+1)^*}, \mathrm{w}_j, (-\psi)^{(d)}) / \sum_{q=1}^{Q} f(I_j = 0 \mid y_{jq}^{(d+1)^*}, \mathrm{w}_j, (-\psi)^{(d)}), \quad (6.93)$$

for all $q = 1,..., Q$. Under nearest neighbour imputation the $Q/2$ nearest neighbours above and below the predicted value for nonrespondent j are used to obtain the Q possible values for imputation, where the value for Q is an even number. The selection of the imputed value for nonrespondent j is based on the probability (6.93). Note that in each I-step only one value is imputed for each nonrespondent.

To obtain the M sets of multiple imputations as a basis for further inference we do the following. Suppose that the data augmentation algorithm has run long enough to achieve approximate stationarity and to be independent of the initial starting value $\zeta^{(0)}$, i.e. d is large enough such that the vector of parameters $\zeta^{(d)}$ are essentially draws from the observed-data posterior. The initial period until convergence of the algorithm is called the burn-in period. To obtain the M sets of multiple imputations it is not possible to use successive sets of imputed values $\tilde{Y}_{mis}$ since they are correlated. It is necessary to either subsample the Markov chain and to use every c-th iterate, where c is chosen large enough to insure independence, or to run M parallel chains of the algorithm until convergence and to take the last set of imputed values from each chain. For computational convenience it was decided to subsample the Markov chain after an initial burn-in period. The results of the M completed datasets are then combined to produce a single overall inference following the rules for multiple imputation given in section 1.4.5, such that we obtain the point estimator, $\hat{\theta}.$, via

$$\hat{\theta}. = \frac{1}{M} \sum_{m=1}^{M} \hat{\theta}.^{(m)} \quad (6.94)$$

and the variance of $\hat{\theta}.$ via

$$\hat{T}. = \bar{G}. + (1 + 1/M)\hat{B}. . \quad (6.95)$$

A heuristic justification for these formulae is given in Schafer (1997) showing that the quantities (6.94) and (6.95) approximate the observed-data posterior moments, the expectation and the variance of the parameter of interest θ. Rubin (1987) uses Bayesian arguments to prove that

$$(\hat{\theta}_{.} - \theta)/\sqrt{\hat{T}_{.}} \sim t_{v}, \tag{6.96}$$

where the degrees of freedom v are given in (1.56), is an approximate observed-data posterior distribution for θ based on the reduced information in $\hat{\theta}_{.}^{(m)}$ and $\bar{G}_{.}^{(m)}$, for $m = 1,...,M$.

e.) Drawbacks of Data Augmentation Approach

Methods such as multiple imputation and data augmentation have certain drawbacks that are briefly addressed here. One drawback is that assumptions about the underlying distributions are necessary, since classical data augmentation is fully parametric. These assumptions are difficult to verify and may not hold in reality. The distributional assumptions are strictly speaking untestable since they refer to the complete model. However, the assumptions made might be good approximations to the real data. We try to address this problem by introducing hot deck imputation in the *I*-step to relax the dependency on distributional assumptions. Furthermore, since the method is developed in a Bayesian framework prior distributions need to be specified and certain assumptions about these distributions are necessary. The specification of uninformative priors avoids making strong assumptions about the priors and often reduces the estimation problem to the problem of maximizing the complete-data likelihood. Note that even the use of different noninformative priors can have a non-negligible impact on the results. In addition, multiple imputation assumes that the rows of the data matrix are independent. It is therefore difficult to incorporate complex sampling designs such as stratification and clustering.

6.3.5.4 Simulation Study

a.) Design of the Simulation Study

The aim of this simulation study is to assess the performance of the estimators derived using data augmentation under the CME assumption. The results for data augmentation based on fully parametric random regression imputation are compared with the results based on hot deck

imputation within classes and nearest neighbour imputation. The method is investigated under misspecification of the response mechanism and the imputation model. In addition, the simulation study evaluates the performance of the point estimators for data augmentation under the assumption of MAR as well as other MAR-based imputation methods. The design of the simulation study is as described in section 3.3, drawing A independent samples $s^{(a)}$, $a = 1,...,A$, with replacement from the original LFS dataset March-May 2000. Since the evaluation of data augmentation in a simulation study is computer intensive the following amendments are made to the design of the study. Note that the results reported here are therefore not directly comparable to the results for the imputation methods reported in chapters 3 to 5. The size of each sample $s^{(a)}$ is reduced to n=1000 units and the number of iterations is limited to A =100 such that 100 independent samples are generated. The models generating the data include fewer variables than in the simulation study in chapters 3 to 5. In the simulation study as well as in the application to LFS data we used $\ln(Y)$ and $\ln(X)$ rather than Y and X. The variable $\ln(Y)$ is generated since it is subject to missing data in the LFS sample and $\ln(X)$ is generated to avoid duplications of units. The models generating $\ln(Y)$ and $\ln(X)$ are as follows:

The variable $\ln(Y)$ is generated using a regression model employing the derived variable $\ln(X)$, $\ln(X)^2$, age, age^2, part-time vs full-time and female adding on a normal error. (6.97)

The regression model for $\ln(X)$ employs age, age^2, part-time vs full-time, female and married, adding on a normal error. (6.98)

Note that the model generating $\ln(X)$ differs slightly from the model generating $\ln(Y)$. The values of the covariates necessary in the simulation study are drawn from the LFS sample using the bootstrap. To generate the indicator variable of response I in sample $s^{(a)}$ following the assumption of CME the probabilities of response are obtained using a logistic regression model based on Y and other covariates similarly as in section 6.2.2. Estimated coefficients necessary for this model to generate the probabilities of response are obtained by fitting a logistic regression model relating I to the variable $\ln(X)$ and other covariates in the original LFS sample. Later on, other nonresponse mechanisms are also considered.

The CME nonresponse mechanism is based on a logistic regression model including the variables $\ln(Y)$, $\ln(Y)^2$, age, age^2, part-time vs full-time and female. (6.99)

Imputation is carried out based on data augmentation under the CME assumption using $\ln(Y)$ and $\ln(X)$ applying random regression, nearest neighbour and hot deck imputation within classes. The procedure is as follows, given a current estimate of the parameters $\zeta^{(d)} = (\zeta_1^{(d)\prime}, \psi^{(d)\prime})'$:

I-step: $\ln(y_j^{(d+1)}) \sim f(\ln(y_j) \mid \ln(x_j), I_j = 0, \mathrm{w}_j, \zeta_{1,mis}^{(d)})$

$$\propto f(\ln(y_j) \mid \ln(x_j), \mathrm{w}_j, \zeta_1^{(d)}) f(I_j = 0 \mid \ln(y_j), \mathrm{w}_j, (-\psi)^{(d)})$$

1.) draw $\ln(\hat{y}_j^{(d+1)*}) \sim f(\ln(y_j) \mid \ln(x_j), \mathrm{w}_j, \zeta^{(d)})$ for nonrespondent $j \in \bar{r}$

2.) accept this value for imputation with probability

$$f(I_j = 0 \mid \ln(y_j^{(d+1)*}), \mathrm{w}_j, (-\psi)^{(d)}) = 1 - f(I_j = 1 \mid \ln(y_j^{(d+1)*}), \mathrm{w}_j, \psi^{(d)}) = \hat{\varrho}_j^{(d+1)*}.$$

If the value $\ln(\hat{y}_j^{(d+1)*})$ is accepted for imputation set $\ln(\hat{y}_j^{(d+1)*}) = \ln(\hat{y}_j^{(d+1)})$, where $\ln(\hat{y}_j^{(d+1)})$ is the imputed value for the nonrespondent j in iteration $d+1$. If the value is rejected return to step 1.).

P-step:

$$\zeta_1^{(d+1)} = (\beta^{(d+1)\prime}, \sigma^2_{Y|X,W}{}^{(d+1)})' \sim f(\zeta_1 \mid (\ln(\tilde{Y}_{obs})', \ln(\tilde{Y}_{mis}^{(d+1)})', \tilde{X}, \tilde{W}), \text{ and}$$

$$\psi^{(d+1)} \sim f(\psi \mid (\ln(\tilde{Y}_{obs})', \ln(\tilde{Y}_{mis}^{(d+1)})', \tilde{I}, \tilde{W}) \tag{6.100}$$

where $\sigma^2_{Y|X,W}{}^{(d+1)} \mid (\ln(\tilde{Y}_{obs})', \ln(\tilde{Y}_{mis}^{(d+1)})', \tilde{W} \sim \hat{\tilde{\varepsilon}}_Y' \hat{\tilde{\varepsilon}}_Y \chi^{-2}_{n-1}$,

$\beta^{(d+1)} \mid \sigma^2_{Y|X,W}{}^{(d+1)}, (\ln(\tilde{Y}_{obs})', \ln(\tilde{Y}_{mis}^{(d+1)})', \tilde{W} \sim N(\hat{\beta}, \sigma^2_{Y|X,W}{}^{(d+1)} (\tilde{\eta}'\tilde{\eta})^{-1})$

and $\psi^{(d+1)} \sim N(\hat{\psi}, \tilde{B}^{-1})$.

The specifications for the data augmentation procedure are as follows. The generation of a single Markov chain is used as opposed to the generation of M parallel chains. In each iteration d, $d=1,...,D$ one value is imputed for each nonrespondent. For nearest neighbour imputation and hot deck imputation within classes the value Q determining the acceptance-rejection sampling algorithm is chosen to be 10 and 20. After convergence of the algorithm M independent sets of imputed values are selected. Discarding the values from the initial burn-in period of length D^*, the chain is subsampled with subsampling constant c to obtain M independent sets of imputed values such that we have $\ln(\tilde{Y}_{mis}^{D^*}), \ln(\tilde{Y}_{mis}^{D^*+c}), \ln(\tilde{Y}_{mis}^{D^*+2c}), ..., \ln(\tilde{Y}_{mis}^{D^*+(M-1)c})$. The constant c needs to be determined such that the subsamples are independent, which will be discussed in section b.) about the assessment of convergence. The number of multiple imputations are chosen to be M=10. The initial starting values for the parameters $\zeta^{(0)}$ required in the I-step are obtained using maximum likelihood estimates based on the observed data in sample $s^{(a)}$, i.e. the model for $\ln(Y)$ conditioning on $\ln(X)$ and W and the model for the nonresponse are fitted based on the respondents in $s^{(a)}$. Note that for this purpose the nonresponse model in the I-step is fitted to observed data with the variable $\ln(X)$ instead of $\ln(Y)$. The covariates defined as W in the theory of data augmentation are chosen in the way that under ideal conditions the variables used in the models required for the I-step coincide with the model generating $\ln(Y)$ and the model generating the nonresponse mechanism. The covariates W are: age, age^2, part-time vs full-time and female. The model for the nonresponse in the I-step is called the nonresponse model whereas the model generating the nonresponse in $s^{(a)}$ is referred to as the nonresponse mechanism.

The variables in the imputation model in the I-step are $\ln(X)$, $\ln(X)^2$, age, age^2, part-time vs full-time and female. (6.101)

The variables in the nonresponse model in the I-step are $\ln(Y)$, $\ln(Y)^2$, age, age^2, part-time vs full-time and female. (6.102)

To facilitate our discussion the following abbreviations are used. Data augmentation based on the assumption of CME using either the weighted bootstrap or rejection sampling in the I-step is referred to as DA-CME, whereas DA-MAR refers to data augmentation under the assumption of MAR without rejection sampling. The abbreviations NN, HDI and reg imp refer to nearest neighbour imputation, hot deck imputation within classes and random regression imputation

respectively. We refer to MAR or CME nonresponse depending on whether the nonresponse mechanism follows the MAR assumption or the CME assumption.

The performance of the following three estimators is investigated: $\hat{P}_1.$, the proportion of employees paid below the NMW of £3.60, $\hat{P}_2.$ the proportion of employees paid between £3.60 and £5, and $\hat{P}_3.$, the proportion of employees earning at the NMW, i.e. between £3.59 and £3.61. The true proportions in the simulation study are: P_1 =0.95%, P_2 =27.41% and P_3 =10.41%. These values are obtained using the average of these quantitities based on complete data within each iteration in the same way as done in section 3.3, equation (3.28). No distinction between different age groups is made.

b.) Assessment of Convergence of Data Augmentation under CME

Before discussing the results for the point estimators the convergence behaviour of the data augmentation algorithm is investigated. To assess the convergence time series plots are used plotting iterates of components of ζ from a single run. Time series plots over the first 200 iterations as well as the first 1000 iterations are shown in the appendix A6.5 for data augmentation under CME using random regression imputation and in appendix A6.6 using nearest neighbour imputation. All components of ζ were analysed. However, only the plots of the coefficient for the derived variable, the variable age and the standard deviation from the imputation model are shown. Examining these plots we conclude that a burn-in period of 200 iterates seems sufficient to achieve convergence. After the burn-in period M independent subsamples $\ln(Y_{mis}^{D^*}), \ln(Y_{mis}^{D^*+c}), \ln(Y_{mis}^{D^*+2c}), ..., \ln(Y_{mis}^{D^*+(M-1)c})$ of the Markov chain are obtained. To examine the relationship among successive iterates and to determine the subsampling constant c sample autocorrelation functions (ACF) are computed. Let $\lambda = \lambda(\zeta)$ denote an individual component or function of the vector of parameters ζ. The lag-c autocorrelation for a stationary series $\{\lambda^{(d)} : d = 1,2,...\}$ is defined as

$$\tau_c = \frac{\mathrm{cov}(\lambda^{(d)}, \lambda^{(d+c)})}{V(\lambda^{(d)})} \tag{6.103}$$

and a sample estimate of τ_c is given by

$$r_c = \frac{\sum_{d=1}^{D-c}(\lambda^{(d)} - \bar{\lambda})(\lambda^{(d+c)} - \bar{\lambda})}{\sum_{d=1}^{D}(\lambda^{(d)} - \bar{\lambda})^2}, \tag{6.104}$$

where $\bar{\lambda} = \sum_{d=1}^{D} \lambda^{(d)}$ is the mean of the series.

Sample autocorrelation plots are used, called ACF plots, plotting r_c versus c for certain values of c, such as $c = 1,...,100$. These provide a useful summary of serial dependence. To calculate r_c D=1100 data augmentation iterations are used. To be able to evaluate if the autocorrelation is significant or not it is useful to carry out the following test statistic (Schafer, 1997). For a stationary normal process that dies out after lag c', i.e. $\tau_c = 0$ for all $c > c'$, the variance of r_c for $c > c'$ is approximately

$$V(r_c) \doteq \frac{1}{D}(1 + 2\sum_{d=1}^{c'} r_d^2), \tag{6.105}$$

where D is the length of the series. The distribution of r_c is approximately normal when $\tau_c = 0$. Therefore we can use the following α-level test statistic. The null hypothesis states that there is no correlation at lag c or beyond, i.e.

$$\tau_c = \tau_{c+1} = \tau_{c+2} = ... = 0, \tag{6.106}$$

versus the alternative hypothesis $\tau_c \neq 0$. The null hypothesis is rejected if

$$|r_c| \geq z_{1-\alpha/2}\left[\frac{1}{D}(1 + 2\sum_{d=1}^{c-1} r_d^2)\right]^{1/2}. \tag{6.107}$$

The critical values for a 0.05-level test are included in the ACF plots as dashed lines. Selected sample ACF plots are given in the appendix A6.7 for the coefficients of the derived variable, age and the standard deviation from the predictive distribution of $\ln(Y)$. The plots show that the serial dependence reduces very quickly after a few iterations. We conclude that c=100 seems sufficient to be able to obtain independent subsamples of the chain. Note that time series and ACF plots only give an indication of the convergence behaviour and may not be foolproof. Convergence with respect to some components of the parameters does not necessarily imply

convergence in the global sense. Schafer (1997) recommends the analysis of scalar functions of ζ for which convergence is likely to be slow. Convergence with respect to this function strengthens the evidence of global convergence. The convergence behaviour of a linear function of the form $b\zeta$ for some constant row-vector b is analysed, where the vector b can be found via the convergence behaviour of the EM-algorithm. However, this method has not been adopted here. Further discussion of the analysis of convergence are given in Schafer (1997), Tanner (1996), Gilks, Richardson and Spiegelhalter (1996), Rosenthal (1993) and Cowles and Bradley (1996).

c.) Performance of Data Augmentation Under CME Under Ideal Model Conditions

The results for data augmentation based on the CME assumption under ideal model conditions are presented. Here, the imputation model includes the same variables as the model generating $\ln(Y)$ in $s^{(a)}$ and the nonresponse model includes the same variables as the model generating the nonresponse mechanism in $s^{(a)}$. The models involved are as specified in (6.97) to (6.102). Based on the analysis of convergence the burn-in time is 200, the subsampling constant c is 100 and the resulting total number of iterations is D=1100 for each sample $s^{(a)}$. M=10 independent sets of imputed values are selected. The results for data augmentation under the CME assumption using random regression, nearest neighbour and hot deck imputation within classes under ideal conditions are presented in table 6.19. Under ideal conditions the fully parametric approach based on random regression imputation with rejection sampling performs well with all three point estimators being approximately unbiased. Also under hot deck imputation within classes all three point estimators are not significantly biased. However, we observe a higher relative bias for $\hat{P}_1.$, which is assumed to be related to the choice of classes. The results reported here refer to hot deck imputation based on nine imputation classes. However, the results for this method may depend on the choice of classes in particular on classes defined for the bottom end of the distribution of the predicted values of Y. Hot deck imputation with only seven classes was also used and the particular choice of classes caused an increase in the relative bias of the estimator $\hat{P}_1..$ For this choice of classes $\hat{P}_1.$ was significantly biased with a relative bias of 13%, whereas both other estimators were approximately unbiased. (The results are not reported here). We conclude that in this simulation study there is evidence that depending on the choice of classes hot deck imputation within classes may lead to biased results.

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

Table 6.19: Bias and relative bias of the three point estimators for data augmentation based on the assumption of CME using random regression, nearest neighbour and hot deck imputation within classes and under ideal model conditions and CME nonresponse. The burn-in period is 200 iterations and the subsampling constant c is 100. (A star (*) indicates that the bias is significantly different from zero on a 95% significance level.)

The results under nearest neighbour imputation are not significantly biased for $\hat{P}_1$. and $\hat{P}_3$.. The bias for $\hat{P}_2$. is significant, however, the relative bias is less than 2%. This is in contrast to earlier findings in chapter 5 where under the assumption of MAR all results for nearest neighbour imputation are approximately unbiased. The bias for nearest neighbour imputation under the CME assumption might be caused by the use of the weighted bootstrap method, which only allows for an approximation to the distribution of the nonrespondents. Ideally rejection sampling in the same way as under random regression imputation should be used, such that a large number of imputed values can be generated which are either accepted or rejected. However, since under nearest neighbour imputation and hot deck imputation within classes only a relative small number of donors can be identified for each nonrespondent, because of the fact that the donor values should be 'close' neighbours, a weighted bootstrap method needs to be used. The performance of the weighted bootstrap method depends on the value of Q. To improve the results for nearest neighbour and hot deck imputation based on the weighted bootstrap the value for Q is increased to Q=20, since if Q increases the approximation to the distribution of the nonrespondents

improves. For both imputation methods, as expected, we can see a reduction in the bias, apart from NN imputation for the estimator $\hat{P}_1$.. However, the improvement is small, such that Q=10 may already be sufficient. The fact that in this simulation study data augmentation based on random regression imputation seems to perform slightly better than the two hot deck methods is thought to be related to using rejection sampling in the *I*-step instead of a weighted bootstrap method.

DA-CME	Bias of $\hat{P}_1$.	Rel. Bias of $\hat{P}_1$.	Bias of $\hat{P}_2$.	Rel. Bias of $\hat{P}_2$.	Bias of $\hat{P}_3$.	Rel. Bias of $\hat{P}_3$.
DA reg imp	$0.70*10^{-3}$ ($1.01*10^{-3}$)	0.62 %	$-0.44*10^{-3}$ ($1.08*10^{-3}$)	-0.26 %	$0.11*10^{-3}$ ($0.18*10^{-3}$)	1.94 %
DA NN, Q=10	$-0.18*10^{-3}$ ($0.33*10^{-3}$)	-1.13 %	$5.53*10^{-3}$ ($1.47*10^{-3}$)*	2.08 %	$1.26*10^{-3}$ ($1.17*10^{-3}$)	1.21 %
DA HDI, Q=10	$0.42*10^{-3}$ ($0.35*10^{-3}$)	4.44 %	$2.94*10^{-3}$ ($1.53*10^{-3}$)	1.07 %	$-0.61*10^{-3}$ ($1.15*10^{-3}$)	-0.58 %

Table 6.20: Bias and relative bias of the three point estimators for data augmentation based on the assumption of CME using random regression, nearest neighbour and hot deck imputation within classes under ideal model conditions and CME nonresponse. The burn-in period is 10 iterations and the subsampling constant *c* is 10.

To analyse the effect of the choice of the burn-in period and the subsampling constant *c* for the data augmentation algorithm shorter burn-in times and smaller subsampling constants *c* are used. The results for a burn-in time of D^* =10 iterations and a subsampling constant *c*=10 are reported in table 6.20. The results are very similar to the results in table 6.19 for a burn-in period of 200 iterations and a subsampling constant *c*=100. We can conclude that a burn-in period of 200 and a subsampling constant of $c = 100$ are therefore sufficient.

To investigate the efficiency of the point estimators under DA-CME the simulation variances are compared. The simulation variance is defined as

$$V(\hat{P}_\cdot) = \frac{1}{A-1}\sum_{a=1}^{A}(\hat{P}_\cdot^{(a)} - \bar{P})^2, \text{ where } \bar{P} = \frac{1}{A}\sum_{a=1}^{A}\hat{P}_\cdot^{(a)}. \qquad (6.108)$$

Table 6.21 shows the simulation variances of the three point estimators. The estimators obtained under data augmentation based on CME are most efficient for random regression imputation. For the two hot deck methods the estimators are slightly more efficient under nearest neighbour imputation than under hot deck imputation within classes, which coincides with findings in chapter 5. There, under the assumption of MAR nearest neighbour imputation was found to be more efficient than hot deck imputation within classes. The simulation variances of data augmentation under the assumption of MAR are given in table 6.34. In comparison to these results the simulation variances for data augmentation under CME are slightly higher. This is expected since for DA-CME additional variability is introduced due to the use of rejection sampling and the weighted bootstrap.

DA-CME	$V(\hat{P}_1.)$	$V(\hat{P}_2.)$	$V(\hat{P}_3.)$
DA reg imp	$0.087*10^{-4}$	$1.796*10^{-4}$	$0.894*10^{-4}$
DA NN, Q=10	$0.118*10^{-4}$	$2.243*10^{-4}$	$1.289*10^{-4}$
DA NN, Q=20	$0.117*10^{-4}$	$2.164*10^{-4}$	$1.241*10^{-4}$
DA HDI, Q=10	$0.127*10^{-4}$	$2.355*10^{-4}$	$1.432*10^{-4}$
DA HDI, Q=20	$0.130*10^{-4}$	$2.359*10^{-4}$	$1.434*10^{-4}$

Table 6.21: Simulation variance of the three point estimators for data augmentation under the assumption of CME using random regression imputation, nearest neighbour imputation and hot deck imputation under CME nonresponse.

We also analyse the performance of the multiple imputation variance estimator, given in section 1.4.5. The variance formula given in (6.95) is denoted $\hat{\text{var}}_{MI}(\hat{P}.)$. Bias and relative bias of the variance estimator are given in table 6.22, the coverage rates are given in table 6.23. The results are based on $A = 100$ iterations of the simulation study in the same way as the results for the point estimators. We found, however, that to obtain more reliable results for the variance estimators, in particular for the coverage rates, it would be preferable to use $A = 1000$ iterations or more, which, because of computing time, has not been done here. To evaluate the significance of the (estimated) bias of the variance estimator the standard error is given in brackets, defined as

$$ste(\hat{bias}(\hat{V})) = \sqrt{\frac{V(\hat{V})}{A}} = \sqrt{\frac{1}{A(A-1)}\sum_{a=1}^{A}(\hat{V}^{(a)} - \bar{V})^2}, \qquad (6.109)$$

where $\bar{V} = \frac{1}{A}\sum_{a=1}^{A}\hat{V}^{(a)}$.

DA-CME	Bias $\hat{var}_{MI}(\hat{P}_1.)$	Bias $\hat{var}_{MI}(\hat{P}_2.)$	Bias $\hat{var}_{MI}(\hat{P}_3.)$	Rel. Bias $\hat{var}_{MI}(\hat{P}_1.)$	Rel. Bias $\hat{var}_{MI}(\hat{P}_2.)$	Rel. Bias $\hat{var}_{MI}(\hat{P}_3.)$
DA reg imp	-0.11*10^{-6} (0.10*10^{-6})	3.07*10^{-6} (1.71*10^{-6})	-1.93*10^{-6} (1.88*10^{-6})	-1.27 %	3.44 %	-2.16 %
DA NN, Q=10	-0.73*10^{-6} (0.64*10^{-6})	13.71*10^{-6} (6.88*10^{-6})*	-4.44*10^{-6} (3.18*10^{-6})	-6.11 %	6.15 %	-3.45 %
DA HDI, Q=10	-0.44*10^{-6} (0.45*10^{-6})	12.96*10^{-6} (6.32*10^{-6})*	-7.63*10^{-6} (4.02*10^{-6})	-3.45 %	5.39 %	-5.34 %

Table 6.22: Bias and relative bias of the MI variance estimator $\hat{var}_{MI}(\hat{P}.)$ of the three point estimators for data augmentation under the assumption of CME using random regression imputation, nearest neighbour imputation and hot deck imputation under CME nonresponse. (A (*) indicates significant bias on a 95% significance level).

DA-CME	Coverage $\hat{var}_{MI}(\hat{P}_1.)$	Coverage $\hat{var}_{MI}(\hat{P}_2.)$	Coverage $\hat{var}_{MI}(\hat{P}_3.)$
DA reg imp	94 %	95 %	94 %
DA NN, Q=10	92 %	96 %	95%
DA HDI, Q=10	93 %	95 %	94 %

Table 6.23: Coverage rates for the 95% confidence interval based on the MI variance estimator $\hat{var}_{MI}(\hat{P}.)$ for the three point estimators for data augmentation under the assumption of CME using random regression imputation, nearest neighbour imputation and hot deck imputation under CME nonresponse.

We can see from table 6.22 that under DA-CME regression imputation the variance estimator is approximately unbiased for all three point estimators. The variance estimator is significantly biased only for DA-CME nearest neighbour and hot-deck imputation for $\hat{\text{var}}_{MI}(\hat{P}_2.)$, overestimating the true variance slightly. For all other estimators there is an indication that the true variance is slightly underestimated. The coverage rates are presented in table 6.23, indicating coverage rates close to 95% for the 95% confidence intervals. For $\hat{\text{var}}_{MI}(\hat{P}_1.)$ under nearest neighbour and hot deck imputation the coverage rate is lower than expected with 92% and 93% respectively. We conclude that overall the MI variance estimator $\hat{\text{var}}_{MI}(\hat{P}.)$ seems to estimate the simulation variance reasonably well and is an adequate variance estimator for $\hat{P}..$

For comparison, the naive variance estimator of a proportion applied to imputed data,

$$\hat{\text{var}}_{naive}(\hat{P}.) = \frac{\hat{P}.(1-\hat{P}.)}{n-1}, \tag{6.110}$$

is also computed. Table 6.24 presents the results for the (estimated) bias and the relative bias of the naive variance estimator. For all three imputation methods and all three point estimators the simulation variance is, as expected, considerably underestimated, indicated by a significant negative bias of $\hat{\text{var}}_{naive}(\hat{P}.)$. Also the coverage rates for the 95% confidence intervals are below 95% (table 6.25). We conclude that the naive variance estimator is not an adequate variance estimator in the presence of imputed data using data augmentation, which coincides with findings in chapter 4.

DA-CME	Bias $\hat{var}_{naive}(\hat{P}_1.)$	Bias $\hat{var}_{naive}(\hat{P}_2.)$	Bias $\hat{var}_{naive}(\hat{P}_3.)$	Rel. Bias $\hat{var}_{naive}(\hat{P}_1.)$	Rel. Bias $\hat{var}_{naive}(\hat{P}_2.)$	Rel. Bias $\hat{var}_{naive}(\hat{P}_3.)$
DA reg imp	$-1.63*10^{-6}$ $(0.32*10^{-6})^*$	$-14.92*10^{-6}$ $(0.61*10^{-6})^*$	$-7.88*10^{-6}$ $(0.76*10^{-6})^*$	-18.77 %	-8.31 %	-4.39 %
DA NN	$-2.52*10^{-6}$ $(0.33*10^{-6})^*$	$-23.36*10^{-6}$ $(0.66*10^{-6})^*$	$-34.84*10^{-6}$ $(0.89*10^{-6})^*$	-21.29 %	-10.41 %	-11.75 %
DA HDI	$-2.86*10^{-6}$ $(0.34*10^{-6})^*$	$-35.61*10^{-6}$ $(0.68*10^{-6})^*$	$-50.68*10^{-6}$ $(0.94*10^{-6})^*$	-22.34 %	-15.11 %	-35.39 %

Table 6.24: Bias and relative bias of the naive variance estimator $\hat{var}_{naive}(\hat{P}.)$ of the three point estimators for data augmentation under the assumption of CME using random regression imputation, nearest neighbour imputation and hot deck imputation under CME nonresponse.

DA-CME	Coverage $\hat{var}_{naive}(\hat{P}_1.)$	Coverage $\hat{var}_{naive}(\hat{P}_2.)$	Coverage $\hat{var}_{naive}(\hat{P}_3.)$
DA reg imp	90 %	94%	89 %
DA NN, Q=10	89 %	94 %	88 %
DA HDI, Q=10	91 %	95 %	88 %

Table 6.25: Coverage rates for the 95% confidence interval based on the naive variance estimator $\hat{var}_{naive}(\hat{P}.)$ for the three point estimators for data augmentation under the assumption of CME using random regression imputation, nearest neighbour imputation and hot deck imputation under CME nonresponse.

d.) Performance of Data Augmentation Under CME Under Misspecification of Models

Of particular interest is the performance of the method DA-CME under misspecification of the imputation model and the nonresponse model. First, the performance under misspecification of the imputation model is discussed. To reflect misspecification of the imputation model the model

for Y in the I-step, $f(\ln(Y)|\ln(X),W)$ is specified as in (6.101), whereas the model generating the complete variable $\ln(Y)$ in the simulation study is modified. In addition to the variables in the models generating $\ln(Y)$ and $\ln(X)$ specified in (6.97) and (6.98) more variables are employed in the following way:

Misspecification 1: model (6.97) and (6.98) plus SOC

Misspecification 2: model (6.97) and (6.98) plus SOC, Q and IND

Misspecification 3: model (6.97) and (6.98) plus SOC, Q, IND, EMPMON, SIZE, REG,

where SOC is occupation group, Q qualification, IND industry sector, EMPMON length of time continuously employed, SIZE number of employees at workplace and REG region. Misspecification 4 allows for the fact that the model for $\ln(Y)$ differs from the model for $\ln(X)$, such that the model generating $\ln(X)$ includes in addition to the variables in (6.98) also the variable SOC, whereas the model generating $\ln(Y)$ includes a larger set of variables, the variables as in (6.97) plus SOC, Q, IND, EMPMON, SIZE and REG. All four cases reflect misspecification of the imputation model of the I-step. In addition, we analyse the case where the variable $\ln(X)$ is not generated but taken from the original LFS sample using the bootstrap and the variable $\ln(Y)$ is generated as in (6.97). This has the advantage that more realistic values of $\ln(X)$ are incorporated but has the effect that some units are replicated in sample $s^{(b)}$. Tables 6.26-6.28 give the results under model misspecification of the imputation model for DA-CME reg imp, DA-CME NN and DA-CME HDI respectively. Under data augmentation based on the CME assumption using random regression imputation (table 6.26) the point estimators $\hat{P}_{1.}$ and $\hat{P}_{3.}$ are not significantly biased under all cases of misspecification. However, with an increasing level of misspecification (misspecification 1-4) the bias increases for all three point estimators. The relative bias of $\hat{P}_{1.}$ is increased from before 1% to now between 2%-8%. The point estimator $\hat{P}_{2.}$ is significantly biased for misspecifications 1-4. However, the relative bias is less than 2%. In the case where the variable $\ln(X)$ is not generated all point estimators are approximately unbiased. We conclude that DA-CME under random regression imputation shows an effect of model misspecifications but overall seems reasonably robust against the model misspecifications considered here. This is slightly surprising since parametric random regression imputation depends on the specifications of the underlying model and has been found to be sensitive to model misspecifications (Schenker and Taylor, 1996). One explanation is that the imputation model is

only used to generate potential values of $\ln(Y)$ for imputation and that rejection sampling is performed accepting or rejecting these values for imputation, such that the correct specification of the nonresponse model may be of greater importance than the specification of the imputation model. As long as the model for the nonresponse used in the acceptance-rejection sampling algorithm is correctly specified the method is expected to work well.

DA-CME, reg imp	Bias of $\hat{P}_1$.	Rel. Bias of $\hat{P}_1$.	Bias of $\hat{P}_2$.	Rel. Bias of $\hat{P}_2$.	Bias of $\hat{P}_3$.	Rel. Bias of $\hat{P}_3$.
Misspecification 1	$0.25*10^{-3}$ ($0.23*10^{-3}$)	4.27 %	$-3.28*10^{-3}$ ($1.44*10^{-3}$)*	-1.22 %	$-1.68*10^{-3}$ ($0.90*10^{-3}$)	-0.17 %
Misspecification 2	$0.31*10^{-3}$ ($0.21*10^{-3}$)	5.70 %	$-3.73*10^{-3}$ ($1.41*10^{-3}$)*	-1.39 %	$-0.18*10^{-3}$ ($0.82*10^{-3}$)	-0.19 %
Misspecification 3	$0.39*10^{-3}$ ($0.21*10^{-3}$)	8.22 %	$-3.73*10^{-3}$ ($1.34*10^{-3}$)*	-1.39 %	$-0.09*10^{-3}$ ($0.85*10^{-3}$)	-0.10 %
Misspecification 4	$0.32*10^{-3}$ ($0.19*10^{-3}$)	8.08 %	$-4.54*10^{-3}$ ($1.45*10^{-3}$)*	-1.71%	$1.78*10^{-3}$ ($0.80*10^{-3}$)	0.19 %
$\ln(X)$ not generated	$0.09*10^{-3}$ ($0.22*10^{-3}$)	2.49 %	$0.48*10^{-3}$ ($1.34*10^{-3}$)	0.20 %	$-0.09*10^{-3}$ ($0.61*10^{-3}$)	-0.17 %

Table 6.26: Bias and relative bias of the three point estimators for data augmentation based on the assumption of CME using random regression imputation under misspecification of the imputation model. The nonresponse mechanism is CME.

DA-CME, NN, Q=10	Bias of $\hat{P}_1$.	Rel. Bias of $\hat{P}_1$.	Bias of $\hat{P}_2$.	Rel. Bias of $\hat{P}_2$.	Bias of $\hat{P}_3$.	Rel. Bias of $\hat{P}_3$.
Misspecification 1	-0.03*10^{-3} (0.26*10^{-3})	-0.53 %	5.01*10^{-3} (1.83*10^{-3})*	1.90 %	1.20*10^{-3} (1.02*10^{-3})	1.22 %
Misspecification 2	-0.01*10^{-3} (0.25*10^{-3})	-0.16 %	6.08*10^{-3} (1.78*10^{-3})*	2.28 %	1.25*10^{-3} (1.07*10^{-3})	1.34 %
Misspecification 3	0.12*10^{-3} (0.25*10^{-3})	2.54 %	7.28*10^{-3} (1.80*10^{-3})*	2.74 %	2.43*10^{-3} (1.02*10^{-3})*	2.59 %
Misspecification 4	0.06*10^{-3} (0.22*10^{-3})	1.56 %	7.39*10^{-3} (1.71*10^{-3})*	2.78 %	3.48*10^{-3} (1.04*10^{-3})*	3.71 %
$\ln(X)$ not generated	-0.07*10^{-3} (0.21*10^{-3})	-2.01 %	-1.39*10^{-3} (1.55*10^{-3})	-0.59 %	-0.81*10^{-3} (0.75*10^{-3})	-1.61 %

Table 6.27: Bias and relative bias of the three point estimators for data augmentation based on the assumption of CME using nearest neighbour imputation under misspecification of the imputation model. The nonresponse mechanism is CME.

DA-CME, HDI, Q=10	Bias of $\hat{P}_1$.	Rel. Bias of $\hat{P}_1$.	Bias of $\hat{P}_2$.	Rel. Bias of $\hat{P}_2$.	Bias of $\hat{P}_3$.	Rel. Bias of $\hat{P}_3$.
Misspecification 1	0.53*10^{-3} (0.27*10^{-3})	8.88 %	5.65*10^{-3} (1.82*10^{-3})*	2.10 %	1.30*10^{-3} (1.00*10^{-3})	1.33 %
Misspecification 2	0.58*10^{-3} (0.25*10^{-3})*	10.65 %	6.86*10^{-3} (1.66*10^{-3})*	2.55 %	1.53*10^{-3} (1.10*10^{-3})	1.56 %
Misspecification 3	0.47*10^{-3} (0.26*10^{-3})	9.79 %	10.07*10^{-3} (1.90*10^{-3})*	3.74 %	3.52*10^{-3} (1.70*10^{-3})*	3.59 %
Misspecification 4	0.32*10^{-3} (0.23*10^{-3})	8.08 %	10.69*10^{-3} (1.75*10^{-3})*	3.97 %	3.72*10^{-3} (1.30*10^{-3})*	3.79 %
$\ln(X)$ not generated	-0.07*10^{-3} (0.21*10^{-3})	-2.03 %	-0.42*10^{-3} (1.50*10^{-3})	-0.18 %	3.04*10^{-3} (0.75*10^{-3})*	-3.10 %

Table 6.28: Bias and relative bias of the three point estimators for data augmentation based on the assumption of CME using hot deck imputation within classes under misspecification of the imputation model. The nonresponse mechanism is CME.

For DA-CME using nearest neighbour imputation we can see from table 6.27 that with an increasing level of misspecification (misspecification 1-4) the bias increases for all three point estimators. The point estimator $\hat{P}_1.$ is not significantly biased for all misspecifications considered here, whereas the point estimator $\hat{P}_2.$ is significantly biased for all cases of misspecification. However, the relative bias is less than 3%. We observe that for misspecification 3 and 4 also $\hat{P}_3.$ is significantly biased. The relative bias is less than 4%. For the case where $\ln(X)$ is not generated all three point estimators are approximately unbiased. DA-CME using nearest neighbour imputation seems to be reasonably robust against model misspecification. The method seems to be slightly more sensitive to misspecification of the imputation model than DA-CME under random regression imputation. This might be related to the fact that DA-CME based on nearest neighbour imputation uses the weighted bootstrap instead of rejection sampling. Under the rejection sampling procedure and assuming that the nonresponse model in the *I*-step is correctly specified the nonresponse model only accepts 'suitable' values for $\ln(Y)$ even if the underlying imputation model is misspecified. Under the weighted bootstrap, however, values for $\ln(Y)$ generated under misspecification of the imputation model are more likely to be accepted since one out of Q values is selected for imputation instead of rejecting a certain number of values until one value is accepted.

The results for DA-CME under hot deck imputation within classes are given in table 6.28. A greater increase in the bias of the point estimators, particularly for $\hat{P}_1.$, than for the two other methods is observed, which is thought to be related to the choice of classes. For all cases of misspecification $\hat{P}_2.$ is significantly biased and $\hat{P}_3.$ is significantly biased for misspecification 3 and 4 and for the case where $\ln(X)$ is not generated. However, all significant relative biases are less than 4%. There is an indication that DA-CME based on hot deck imputation within classes is more sensitive to misspecification of the imputation model than DA-CME based on random regression imputation, which, similarly to the case of nearest neighbour imputation, may be caused by the use of the weighted bootstrap instead of rejection sampling.

It is also of interest to analyse data augmentation based on the CME assumption under misspecification of the nonresponse mechanism. In the following analysis the nonresponse model

is as specified in (6.102). The method DA-CME is analysed under different nonresponse mechanisms as follows:

1.) uniform nonresponse

2.) CME nonresponse where the model generating nonresponse in $s^{(a)}$ includes a richer set of covariates than the nonresponse model in the *I*-step. This is referred to as CME +6, since the model generating the nonresponse includes in addition to the variables in (6.99) six more variables: occupation (SOC), qualification (Q), industry sector (IND), length of time continuously employed (EMPMON), number of employees at workplace (SIZE) and region (REG).

3.) MAR nonresponse

4.) A nonresponse mechanism that is based on the full nonresponse model, i.e. including the variables $\ln(Y)$, $\ln(X)$ and other covariates as specified in (6.99), such that the nonresponse is neither MAR nor CME.

DA-CME, reg imp	Bias of $\hat{P}_1$.	Rel. Bias of $\hat{P}_1$.	Bias of $\hat{P}_2$.	Rel. Bias of $\hat{P}_2$.	Bias of $\hat{P}_3$.	Rel. Bias of $\hat{P}_3$.
CME +6	$0.14*10^{-3}$ ($0.28*10^{-3}$)	1.47 %	$0.33*10^{-3}$ ($1.25*10^{-3}$)	0.12 %	$0.17*10^{-3}$ ($0.91*10^{-3}$)	0.16 %
MAR	$-0.21*10^{-3}$ ($0.29*10^{-3}$)	-2.20 %	$-23.19*10^{-3}$ ($1.20*10^{-3}$)*	-8.46 %	$-7.14*10^{-3}$ ($0.93*10^{-3}$)*	-6.86 %
full model	$-0.43*10^{-3}$ ($0.28*10^{-3}$)	-4.61 %	$-16.6*10^{-3}$ ($1.44*10^{-3}$)*	-6.07 %	$-6.83*10^{-3}$ ($0.96*10^{-3}$)*	-6.56 %

Table 6.29: Bias and relative bias of the three point estimators for data augmentation based on the assumption of CME using random regression imputation under misspecification of the nonresponse mechanism.

DA-CME, NN, Q=10	Bias of $\hat{P}_1$.	Rel. Bias of $\hat{P}_1$.	Bias of $\hat{P}_2$.	Rel. Bias of $\hat{P}_2$.	Bias of $\hat{P}_3$.	Rel. Bias of $\hat{P}_3$.
CME +6	$-1.19*10^{-3}$ $(0.32*10^{-3})$	-1.25 %	$5.69*10^{-3}$ $(1.48*10^{-3})^*$	2.07 %	$1.64*10^{-3}$ $(1.14*10^{-3})$	1.58 %
MAR	$-0.35*10^{-3}$ $(0.32*10^{-3})$	-3.77 %	$-4.54*10^{-3}$ $(1.45*10^{-3})^*$	-1.65 %	$-1.69*10^{-3}$ $(1.13*10^{-3})$	-1.63 %
full model	$-2.49*10^{-3}$ $(3.26*10^{-3})$	-2.61 %	$7.40*10^{-3}$ $(2.01*10^{-3})^*$	2.70 %	$0.45*10^{-3}$ $(1.37*10^{-3})$	0.43 %

Table 6.30: Bias and relative bias of the three point estimators for data augmentation based on the assumption of CME using nearest neighbour imputation under misspecification of the nonresponse mechanism.

DA-CME, HDI, Q=10	Bias of $\hat{P}_1$.	Rel. Bias of $\hat{P}_1$.	Bias of $\hat{P}_2$.	Rel. Bias of $\hat{P}_2$.	Bias of $\hat{P}_3$.	Rel. Bias of $\hat{P}_3$.
CME +6	$0.26*10^{-3}$ $(0.33*10^{-3})$	2.79 %	$2.67*10^{-3}$ $(1.52*10^{-3})$	0.97 %	$-0.02*10^{-3}$ $(1.17*10^{-3})$	-0.02 %
MAR	$0.04*10^{-3}$ $(0.31*10^{-3})$	0.41 %	$-8.30*10^{-3}$ $(1.39*10^{-3})^*$	-2.88 %	$-3.23*10^{-3}$ $(1.05*10^{-3})^*$	-2.89 %
full model	$0.11*10^{-3}$ $(0.32*10^{-3})$	1.18 %	$3.70*10^{-3}$ $(2.03*10^{-3})$	1.35 %	$-1.16*10^{-3}$ $(1.34*10^{-3})$	-1.11 %

Table 6.31: Bias and relative bias of the three point estimators for data augmentation based on the assumption of CME using hot deck imputation within classes under misspecification of the nonresponse mechanism.

As expected the point estimators under the simplest case of uniform nonresponse are all approximately unbiased for all three methods such that the results are not reported here. Tables 6.29-6.31 give the results under misspecification of the nonresponse mechanism featuring departure from the assumption of CME for DA-CME using random regression imputation, nearest neighbour imputation and hot deck imputation within classes respectively. As shown in table 6.29 for DA-CME regression imputation all three estimators are not significantly biased for

CME +6. However, under the MAR assumption and the full nonresponse model $\hat{P}_{2\cdot}$ and $\hat{P}_{3\cdot}$ are significantly biased with a relative bias between 6%-8%. The point estimator $\hat{P}_{1\cdot}$ is not significantly biased in these two cases, however, it shows a relative bias of 4.6% under the full model. DA-CME using nearest neighbour imputation (table 6.30) is significantly biased only for $\hat{P}_{2\cdot}$ with less than 2% for all three misspecifications. All other relative biases are less than 2%, apart from for $\hat{P}_{1\cdot}$ under MAR and the full nonresponse model. DA-CME under hot deck imputation (table 6.31) is approximately unbiased in all three point estimators for CME +6 and the full model. It shows a significant bias under the MAR assumption for $\hat{P}_{2\cdot}$ and $\hat{P}_{3\cdot}$.. However, the relative bias is less than 3%. Whereas DA-CME based on random regression imputation seems sensitive to some misspecification of the nonresponse model DA-CME using nearest neighbour and hot deck imputation within classes seem reasonably robust against the misspecifications considered here, which is thought to be related to the use of the weighted bootstrap for DA-CME NN and HDI rather than rejection sampling. Since regression imputation is based on rejection sampling which uses the nonresponse model directly for rejecting or accepting possible values for $\ln(Y)$ this method is more sensitive to misspecification of the nonresponse model than DA-CME NN and HDI based on the weighted bootstrap method if the imputation model is correctly specified. The reasonably good performance of DA-CME NN and HDI under misspecifications of the nonresponse model seems to be an advantage of the CME-based methods in comparison to the MAR-based methods which showed a greater relative bias under misspecification of the nonresponse model with approximately 5%-6% relative bias (see section 6.2.2).

DA-full nonresponse model	Bias of $\hat{P}_{1\cdot}$	Rel. Bias of $\hat{P}_{1\cdot}$	Bias of $\hat{P}_{2\cdot}$	Rel. Bias of $\hat{P}_{2\cdot}$	Bias of $\hat{P}_{3\cdot}$	Rel. Bias of $\hat{P}_{3\cdot}$
DA reg imp	$0.69*10^{-3}$ $(0.29*10^{-3})^*$	7.28 %	$15.19*10^{-3}$ $(2.32*10^{-3})^*$	5.54 %	7.08 $(1.27*10^{-3})^*$	6.80 %

Table 6.32: Bias and relative bias of the three point estimators for data augmentation based on the full nonresponse model using random regression imputation under CME nonresponse.

As discussed in section 6.3.5.3 the factorisation of the likelihood (6.63) requires the use of the common measurement error assumption to identify the nonresponse model. However, we are also interested in the performance of data augmentation using the full nonresponse model, i.e. using $f(I_j = 0 \mid y_j, x_j, \mathrm{w}_j)$ instead of $f(I_j = 0 \mid y_j, \mathrm{w}_j)$ in the *I*-step. Table 6.32 gives the results for data augmentation based on random regression imputation with rejection sampling based on the full nonresponse model. The nonresponse mechanism follows the assumption of CME. As expected the algorithm does not converge and the results obtained under this method are significantly biased, since the nonresponse model is not identified. The time series plots in appendix A6.8 show that the algorithm has not converged even after D=10,000 iterations. It is therefore necessary to make an assumption about the nonresponse model in the imputation step, such as MAR or CME. The fact that the algorithm failed to converge may be used generally to detect potential problems with the specifications of the models involved.

e.) Performance of Data Augmentation under the Assumption of MAR

As a comparison the performance of the point estimators is analysed under data augmentation based on the MAR assumption, without rejection sampling and the weighted bootstrap. In the following data augmentation under MAR is briefly described for the example discussed here with the variable Y subject to nonresponse. The *I*-step and the *P*-step of data augmentation under MAR in the example considered here are as follows. Given a current estimate of the parameters $\zeta_1^{(d)} = (\beta^{(d)'}, \sigma^{2\ (d)}_{Y|X,W})'$ we have

I-step: $$\ln(\hat{y}_j^{(d+1)}) \sim f(\ln(y_j^{(d+1)}) \mid \ln(x_j), \mathrm{w}_j, \zeta_1^{(d)}), \tag{6.111}$$

where $\zeta_1^{(d)}$ refers to the vector of parameters of the complete data model and $\ln(\hat{y}_j^{(d+1)})$ are the imputed values in iteration $d+1$.

P-step: $$\zeta_1^{(d+1)} \sim f(\zeta_1 \mid (\ln(\tilde{Y}_{obs})', \ln(\tilde{Y}_{mis}^{(d+1)})')', \tilde{X}, \tilde{W}) \tag{6.112}$$

where

$$\sigma^{2\ (d+1)}_{Y|X,W} \mid (\ln(\tilde{Y}_{obs})', \ln(\tilde{Y}_{mis}^{(d+1)})')', \tilde{W} \sim \hat{\tilde{\varepsilon}}_Y' \hat{\tilde{\varepsilon}}_Y \chi^{-2}_{n-1},$$

$$\beta^{(d+1)} \mid \sigma^{2}_{Y|X,W}{}^{(d+1)}, (\ln(\tilde{Y}_{obs})', \ln(\tilde{Y}_{mis}^{(d+1)})')', I\tilde{W} \sim N(\hat{\beta}, \sigma^{2}_{Y|X,W}{}^{(d+1)}(\tilde{\eta}'\tilde{\eta})^{-1}).$$

Note that under the assumption of MAR and in the case of a monotone missing data structure, as it is given here, we do not necessarily need to use an iterative procedure such as data augmentation to obtain draws from the observed-data posterior. The particular factorization of the likelihood, where the observed-data likelihood factors into complete-data likelihoods whose parameters are distinct, enables us to express the observed-data posterior in a tractable form. In the case of a monotone missing data structure and if the prior density factors into independent densities, then the observed-data posterior distribution also factors into independent posteriors and Bayesian inference is possible without iteration. In this case the observed-data posterior is the product of a multivariate normal and a scaled inverted-chisquare density based on observed data only. The estimates of the parameters from the observed-data posterior are drawn based on observed data only. We have

$$f(\zeta_1 \mid \ln(\tilde{Y}_{obs}), \tilde{X}, I\tilde{W}) \propto (\sigma^{2}_{Y|X,W})^{-1} \exp\left\{\frac{-1}{2\sigma^{2}_{Y|X,W}}(\beta-\hat{\beta})'\tilde{\eta}'_{obs}\tilde{\eta}_{obs}(\beta-\hat{\beta})\right\}$$

$$*(\sigma^{2}_{Y|X,W})^{-(\frac{n-1}{2})-1} \exp\left\{-\frac{1}{2}\hat{\tilde{\varepsilon}}'_{Y}\hat{\tilde{\varepsilon}}_{Y}\sigma^{2}_{Y|X,W}\right\},$$

such that

$$\sigma^{2}_{Y|X,W} \mid \ln(\tilde{Y}_{obs}), I\tilde{W} \sim \hat{\tilde{\varepsilon}}'_{Y}\hat{\tilde{\varepsilon}}_{Y}\chi^{-2}_{n-1} \tag{6.113}$$

$$\beta \mid \sigma^{2}_{Y|X,W}, \ln(\tilde{Y}_{obs}), I\tilde{W} \sim N(\hat{\beta}, \sigma^{2}_{Y|X,W}(\tilde{\eta}'_{obs}\tilde{\eta}_{obs})^{-1}), \tag{6.114}$$

where β and η are defined in (6.66), $\tilde{\eta}_{obs}$ refers to respondent data-matrix only, $\hat{\beta}$ is the ML estimate based on respondents data only, $\hat{\beta} = (\tilde{\eta}'_{obs}\tilde{\eta}_{obs})^{-1}\tilde{\eta}'_{obs}\ln(\tilde{Y}_{obs})$ and $\hat{\tilde{\varepsilon}}_{Y} = \ln(\tilde{Y}_{obs}) - \tilde{\eta}_{obs}\hat{\beta}$ based on respondents only instead of the augmented data. Because of the tractability of the observed-data posterior the data augmentation iteration converges immediately if the initial starting value of the vector of parameters ζ_1 in the data augmentation procedure is the maximum likelihood estimate based on observed data. For other examples of this kind see Schafer (1997, p. 18 and p.

73). In general, however, the observed-data posterior is not tractable and iterative methods such as data augmentation need to be used.

Here, under the assumption of MAR we have used data augmentation instead of direct draws from the observed-data posterior. The burn-in period is 200 iterations and the subsampling constant c is 100 as before. Selected time series plots and sample autocorrelation function plots are given in the appendix A6.9 and A6.10 respectively. We can see that, as expected, these specifications are sufficient for the convergence of the algorithm. The results of the three point estimators for data augmentation based on the MAR assumption, referred to as DA-MAR, using random regression imputation, nearest neighbour and hot deck imputation within classes are presented in table 6.33. For DA-MAR using random regression imputation and nearest neighbour imputation all point estimators are approximately unbiased under ideal model conditions, i.e. the imputation model coincides with the model generating $\ln(Y)$. For DA-MAR using hot deck imputation only $\hat{P}_1$. is significantly biased, which is thought to be related to the choice of classes since particularly at the bottom end of the distribution of the predicted values for $\ln(Y)$ it may be difficult to define the classes adequately. Overall, DA-MAR based on random regression imputation, nearest neighbour and hot deck imputation performs well.

DA-MAR	Bias of $\hat{P}_1$.	Rel. Bias of $\hat{P}_1$.	Bias of $\hat{P}_2$.	Rel. Bias of $\hat{P}_2$.	Bias of $\hat{P}_3$.	Rel. Bias of $\hat{P}_3$.
DA reg imp	$0.01*10^{-3}$ ($0.29*10^{-3}$)	0.07 %	$-0.50*10^{-3}$ ($1.28*10^{-3}$)	-0.18 %	$0.32*10^{-3}$ ($0.94*10^{-3}$)	0.31 %
DA NN	$-0.05*10^{-3}$ ($0.30*10^{-3}$)	-0.54 %	$-0.35*10^{-3}$ ($1.47*10^{-3}$)	-0.12 %	$0.12*10^{-3}$ ($0.10*10^{-3}$)	0.12 %
DA HDI	$2.68*10^{-3}$ ($1.15*10^{-3}$)*	2.37 %	$0.74*10^{-3}$ ($1.47*10^{-3}$)	0.27	$1.46*10^{-3}$ ($1.10*10^{-3}$)	1.40 %

Table 6.33: Bias and relative bias of the three point estimators for data augmentation based on the assumption of MAR using random regression, nearest neighbour and hot deck imputation within classes under MAR nonresponse.

Table 6.34 shows the simulation variances of the three point estimators. As before for DA-CME (see table 6.21) random regression imputation is most efficient. For the two hot deck methods nearest neighbour imputation is slightly more efficient than hot deck imputation within classes. In comparison to the results for DA-CME the simulation variances for DA-MAR presented here are slightly lower. This is expected since for DA-CME additional variability is introduced due to the use of rejection sampling and the weighted bootstrap.

DA-MAR	$V(\hat{P}_1.)$	$V(\hat{P}_2.)$	$V(\hat{P}_3.)$
DA reg imp	$0.088*10^{-4}$	$1.654*10^{-4}$	$0.902*10^{-4}$
DA NN	$0.094*10^{-4}$	$2.185*10^{-4}$	$1.159*10^{-4}$
DA HDI	$0.119*10^{-4}$	$2.182*10^{-4}$	$1.218*10^{-4}$

Table 6.34: Simulation variance of the three point estimators for data augmentation under the assumption of MAR using random regression imputation, nearest neighbour imputation and hot deck imputation under MAR nonresponse.

DA-MAR	Bias $\hat{var}_{MI}(\hat{P}_1.)$	Bias $\hat{var}_{MI}(\hat{P}_2.)$	Bias $\hat{var}_{MI}(\hat{P}_3.)$	Rel. Bias $\hat{var}_{MI}(\hat{P}_1.)$	Rel. Bias $\hat{var}_{MI}(\hat{P}_2.)$	Rel. Bias $\hat{var}_{MI}(\hat{P}_3.)$
DA reg imp	$0.01*10^{-6}$ ($0.01*10^{-6}$)	$4.69*10^{-6}$ ($2.80*10^{-6}$)	$-1.19*10^{-6}$ ($0.78*10^{-6}$)	1.05 %	2.84 %	-1.33 %
DA NN	$0.31*10^{-6}$ ($0.30*10^{-6}$)	$0.90*10^{-6}$ ($0.89*10^{-6}$)	$-4.44*10^{-6}$ ($3.18*10^{-6}$)	3.22 %	4.12 %	-2.48 %
DA HDI	$0.36*10^{-6}$ ($0.37*10^{-6}$)	$-0.32*10^{-6}$ ($1.15*10^{-6}$)*	$-5.02*10^{-6}$ ($3.63*10^{-6}$)	3.10 %	-1.49 %	-4.12 %

Table 6.35: Bias and relative bias of the MI variance estimator $\hat{var}_{MI}(\hat{P}.)$ of the three point estimators for data augmentation under the assumption of MAR using random regression imputation, nearest neighbour imputation and hot deck imputation under MAR nonresponse.

DA-MAR	Coverage $\hat{var}_{MI}(\hat{P}_1.)$	Coverage $\hat{var}_{MI}(\hat{P}_2.)$	Coverage $\hat{var}_{MI}(\hat{P}_3.)$
DA reg imp	93 %	96 %	95 %
DA NN	93 %	95 %	95 %
DA HDI	94 %	96 %	94 %

Table 6.36: Coverage rates for the 95% confidence interval based on the MI variance estimator $\hat{var}_{MI}(\hat{P}.)$ for the three point estimators for data augmentation under the assumption of MAR using random regression imputation, nearest neighbour imputation and hot deck imputation under MAR nonresponse.

DA-MAR	Bias $\hat{var}_{naive}(\hat{P}_1.)$	Bias $\hat{var}_{naive}(\hat{P}_2.)$	Bias $\hat{var}_{naive}(\hat{P}_3.)$	Rel. Bias $\hat{var}_{naive}(\hat{P}_1.)$	Rel. Bias $\hat{var}_{naive}(\hat{P}_2.)$	Rel. Bias $\hat{var}_{naive}(\hat{P}_3.)$
DA reg imp	$-0.58*10^{-6}$ $(0.34*10^{-6})^*$	$-12.10*10^{-6}$ $(0.61*10^{-6})^*$	$-13.69*10^{-6}$ $(0.74*10^{-6})$	-6.67 %	- 7.32%	-15.22%
DA NN	$-0.08*10^{-6}$ $(0.30*10^{-6})$	$-19.98*10^{-6}$ $(0.67*10^{-6})^*$	$-22.69*10^{-6}$ $(0.84*10^{-6})^*$	-0.84 %	-9.14 %	-19.57 %
DA HDI	$-1.16*10^{-6}$ $(0.34*10^{-6})^*$	$-19.56*10^{-6}$ $(0.66*10^{-6})^*$	$-28.83*10^{-6}$ $(0.87*10^{-6})^*$	-9.90 %	-8.95 %	-23.42 %

Table 6.37: Bias and relative bias of the naive variance estimator $\hat{var}_{naive}(\hat{P}.)$ of the three point estimators for data augmentation under the assumption of MAR using random regression imputation, nearest neighbour imputation and hot deck imputation under MAR nonresponse.

DA-MAR	Coverage $\hat{var}_{naive}(\hat{P}_1.)$	Coverage $\hat{var}_{naive}(\hat{P}_2.)$	Coverage $\hat{var}_{naive}(\hat{P}_3.)$
DA reg imp	93 %	96 %	93 %
DA NN	91 %	91 %	93 %
DA HDI	91 %	94 %	92 %

Table 6.38: Coverage rates for the 95% confidence interval based on the naive variance estimator $\hat{var}_{naive}(\hat{P}.)$ for the three point estimators for data augmentation under the assumption of MAR using random regression imputation, nearest neighbour imputation and hot deck imputation under MAR nonresponse.

The multiple imputation variance estimator $\hat{var}_{MI}(\hat{P}.)$ is computed and the results for bias and relative bias are presented in table 6.35. All results are approximately unbiased, apart from $\hat{var}_{MI}(\hat{P}_2.)$ under DA hot deck imputation. The coverage rates for the 95% confidence intervals are close to 95% (table 6.36). Overall, the results show, as expected, a good performance of the MI variance estimator for all three imputation methods. For comparison the naive variance estimator $\hat{var}_{naive}(\hat{P}.)$ is computed, which underestimates the simulation variance considerably (table 6.37), however, less so than in the case of DA-CME. Also the coverage rates are below 95% for the naive variance estimator (table 6.38). We conclude that, as expected, the naive variance estimator is not an adequate variance estimator in the presence of imputed data using data augmentation.

f.) Performance of other MAR-based Imputation Methods

For comparison the performances of the point estimators under other MAR-based methods, i.e. random regression imputation, nearest neighbour imputation and hot deck imputation within classes sampling donors by simple random sampling without replacement as discussed in chapters 2-5, are investigated under the specifications of this simulation study. The imputation methods are repeated such that M=10 values are used for imputation. The methods are denoted reg imp10, NN10 and HDIwor10 respectively. The three imputation methods are evaluated under MAR and CME nonresponse. Table 6.39 presents the performance of the three imputation methods under a MAR nonresponse mechanism. The imputation model coincides with the model generating Y in sample $s^{(a)}$. Under these ideal conditions all estimators for all three models are approximately unbiased with relative biases of less than 1%. We can see a higher relative bias under hot deck imputation for estimator $\hat{P}_1.$, which is thought to be related to the choice of classes at the lower end of the pay distribution. Table 6.40 presents the results for the three imputation methods under CME nonresponse. The two point estimators $\hat{P}_2.$ and $\hat{P}_3.$ are significantly biased with around 6-8% for all three imputation methods, which coincides approximately with the findings in section 6.2.2 where predictive mean matching imputation is analysed under CME nonresponse, however under the simulation design as described in chapter 3. As expected the biases for all three point estimators are higher than when using CME methods under CME nonresponse as shown in table 6.19. Table 6.41 shows the simulation variance of the three point estimators under random regression, nearest neighbour and hot deck imputation within classes. For all three point

estimators random regression imputation is most efficient. For the two hot deck methods nearest neighbour imputation is more efficient than hot deck imputation within classes. These results coincide with findings for the simulation variances under the DA-CME methods given in table 6.21 and under DA-MAR given in table 6.34. In comparison to these results the simulation variances for the MAR based methods presented here are slightly lower. This is expected since for data augmentation methods additional variability is introduced due to the variability of the parameters.

The performances of the multiple imputation variance estimator and the naive variance estimator under the MAR based imputation methods, which are, however, not proper multiple imputations, are also analysed. As expected $\hat{var}_{MI}(\hat{P}.)$ leads to underestimation for all three imputation methods and all three point estimators since the multiple imputations are not proper, which has been discussed in chapter 4. The coverage rates are found to be around or below 95% for the 95% confidence intervals. The results for the naive variance estimator indicate as expected underestimation of the simulation variance and coverage rates below 95%. (The results are not presented here). We conclude that the MI variance estimator and the naive variance estimator are not an adequate estimator in the presence of imputed data based on the imputation methods used here.

	Bias of $\hat{P}_1.$	Rel. Bias of $\hat{P}_1.$	Bias of $\hat{P}_2.$	Rel. Bias of $\hat{P}_2.$	Bias of $\hat{P}_3.$	Rel. Bias of $\hat{P}_3.$
Random Reg Imp10	$0.05*10^{-3}$ ($0.29*10^{-3}$)	0.47 %	$-0.93*10^{-3}$ ($1.27*10^{-3}$)	-0.34 %	$0.24*10^{-3}$ ($0.95*10^{-3}$)	0.23 %
NN10	$-0.04*10^{-3}$ ($0.30*10^{-3}$)	-0.39 %	$0.08*10^{-3}$ ($1.46*10^{-3}$)	0.03 %	$0.11*10^{-3}$ ($1.05*10^{-3}$)	0.10 %
HDIwor10	$0.60*10^{-3}$ ($0.32*10^{-3}$)	6.31 %	$1.72*10^{-3}$ ($1.47*10^{-3}$)	0.62 %	$2.37*10^{-3}$ ($1.07*10^{-3}$)	0.22 %

Table 6.39: Bias and relative bias of the three point estimators for random regression imputation, nearest neighbour and hot deck imputation based on M=10 imputed values under MAR nonresponse.

	Bias of $\hat{P}_1$.	Rel. Bias of $\hat{P}_1$.	Bias of $\hat{P}_2$.	Rel. Bias of $\hat{P}_2$.	Bias of $\hat{P}_3$.	Rel. Bias of $\hat{P}_3$.
Random Reg Imp10	$0.43*10^{-3}$ $(0.29*10^{-3})$	4.61 %	$15.95*10^{-3}$ $(1.35*10^{-3})^*$	5.81 %	$6.86*10^{-3}$ $(0.97*10^{-3})^*$	6.59 %
NN10	$0.30*10^{-3}$ $(0.32*10^{-3})$	3.17 %	$16.57*10^{-3}$ $(1.62*10^{-3})^*$	6.04 %	$6.77*10^{-3}$ $(1.14*10^{-3})^*$	6.51 %
HDIwor10	$1.37*10^{-3}$ $(0.34*10^{-3})^*$	14.44 %	$16.79*10^{-3}$ $(1.59*10^{-3})^*$	6.12 %	$8.18*10^{-3}$ $(1.13*10^{-3})^*$	7.85 %

Table 6.40: Bias and relative bias of the three point estimators for random regression imputation, nearest neighbour and hot deck imputation based on $M=10$ imputed values under CME nonresponse.

	$V(\hat{P}_1.)$	$V(\hat{P}_2.)$	$V(\hat{P}_3.)$
Random Reg Imp10	$0.086*10^{-4}$	$1.631*10^{-4}$	$0.909*10^{-4}$
NN10	$0.094*10^{-4}$	$2.152*10^{-4}$	$1.116*10^{-4}$
HDIwor10	$0.105*10^{-4}$	$2.178*10^{-4}$	$1.162*10^{-4}$

Table 6.41: Simulation variance of the three point estimators for random regression imputation, nearest neighbour imputation and hot deck imputation within classes imputing $M=10$ values, under MAR nonresponse.

6.3.5.5 Application to LFS

In the following data augmentation under the assumption of CME is applied to the March-May 2000 LFS quarter. Again, we used $\ln(Y)$ and $\ln(X)$. The estimates of the three proportions of interest are compared to estimates obtained under MAR-based methods, i.e. random regression imputation, nearest neighbour and hot deck imputation within classes with and without data augmentation. Estimates of proportions of employees that are less well paid obtained for a method valid under the CME assumption may be expected to be lower than for a method valid under the MAR assumption as discussed in section 6.3.4.4. Methods that are valid under the MAR assumption may lead to an overestimate of proportions of employees earning at the lower end of

the pay scale, whereas the CME assumption allows for the fact that the nonresponse depends on true hourly pay.

The specifications for the data augmentation procedures considered here are as follows. The imputation model includes all the variables as specified in table 2.1 in chapter 2. The nonresponse model in the *I*-step includes all these variables and in addition two variables from the nonresponse model as specified in chapter 3 table 3.4. These are 'additions to basic pay' (ADDTBP) and 'proxy response' (PROXY). The data augmentation algorithm is run using a single chain with D=1100, a burn-in period of D^* =200 and a subsampling constant c=100. A shorter chain with D=100, a burn-in period of D^* =10 and a subsampling constant c=10 is also used to analyse the effect of the length of the chain. Similar specifications are applied by Carroll, Ruppert and Stefanski (1995) and Schafer (1997). The initial starting values for the vector of parameters required in the *I*-step are chosen to be the maximum likelihood estimates of the imputation and the nonresponse model based on observed data only. For this purpose the variable $\ln(Y)$ in the nonresponse model is substituted by the variable $\ln(X)$. Also different starting values for the data augmentation algorithm are used to investigate the sensitivity of the algorithm to the choice of the initial starting value. However, setting the initial coefficients for both the imputation as well as the nonresponse model different from the maximum likelihood estimates appears difficult. Simple initial starting values, for example setting all coefficients to 1, 0.1, 0.01, with or without the sign from the maximum likelihood estimates, causes the algorithm to fail in the first one or two iterations. There are several reasons for this. If for example the coefficients for the model $f(\ln(\tilde{Y}) \mid \ln(\tilde{X}), \tilde{W})$ are set in a way that the distribution of the imputed values generated in the *I*-step differs significantly from the observed values for $\ln(Y)$ (for example the imputed values for $\ln(Y)$ are much larger than the observed values for $\ln(Y)$) then the logistic regression model modelling the nonresponse predicts probabilities close to zero and one, which may result in a failure of the data augmentation algorithm. If for example some of the imputed values for $\ln(Y)$ are large the probability of acceptance $\hat{\varrho}_i$ cannot be calculated, since $\hat{\varrho}_i = 1-\{\exp(\tau_i\hat{\psi})/(1+\exp(\tau_i\hat{\psi}))\}$ and $\exp(\tau_i\hat{\psi})$ approaches infinity. If on the other hand the probabilities for acceptance $\hat{\varrho}_i$ are very small, it takes a long time until acceptance of imputed values which causes the algorithm to run for a long time. This means that 'sensible' values need to be chosen as starting values for the data augmentation algorithm for both models, the linear regression model predicting Y and the logistic model

predicting nonresponse. Therefore coefficients are chosen that have the same sign and are of the same order of magnitude as the maximum likelihood estimates to guarantee the generation of sensible imputed values and sensible predicted probabilities of nonresponse. These parameter specifications are referred to as the chosen initial parameter estimates.

The results for data augmentation under the CME assumption are presented in table 6.42. Random regression imputation, nearest neighbour imputation and hot deck imputation within classes are used in the *I*-step. Either rejection sampling or the weighted bootstrap is performed to select values for imputation. Hot deck imputation within classes and nearest neighbour imputation give very similar estimates for all three point estimators. The unweighted proportion of employees earning below the NMW is about 0.44% and the unweighted proportion at the NMW is about 2.35% for employees 18 years and older. For both forms of hot deck imputation, as might be expected, the estimates for P_1 and P_3 under the CME assumption are lower than under the MAR assumption. The reduction for both estimates is approximately 10%. Also the estimate for P_2 obtained under the CME assumption is slightly lower than the estimate under the MAR assumption. The results for $\hat{P}_1$. and $\hat{P}_2$. obtained using random regression imputation based on CME are lower than under the MAR assumption. The estimate for $\hat{P}_3$. is approximately the same. The results under regression imputation differ from the results under hot deck imputation, in the sense that the estimate for P_1 is higher and the estimates for P_2 and P_3 are lower than under the hot deck methods under both CME and MAR. This is assumed to be related to the inadequacy of the use of random regression imputation in the context of hourly earnings data, since the method relies on certain distributional assumptions, such as the assumption of constant variance. It does not preserve truncation and rounding effects well, as present in the distribution of hourly earnings. Hot deck methods are expected to capture these features better and are therefore recommended for practical use. The inadequacy of random regression imputation is illustrated in the application to LFS data without data augmentation at the end of this section. From table 6.42 we can also conclude that the estimates are very similar for different lengths of the chain, with different burn-in periods and subsampling constants *c*, suggesting that the algorithm converges quickly. Selected time series plots and sample autocorrelation function plots for DA-CME under nearest neighbour imputation applied to LFS data are given in the appendix A6.11 and A6.12 respectively. From the time series plots we conclude that the algorithm converges quickly. The autocorrelation plots show

that the subsampling constant c seems to be adequately chosen. From table 6.42 we can also see that the estimates of the proportions of interest produced using initial starting values of the parameters different from the maximum likelihood estimates are very similar to the results under the maximum likelihood estimates. The estimates therefore do not seem to be very dependent on the choice of the initial starting values.

DA-CME	$\hat{P}_1$. in %	$\hat{P}_2$. in %	$\hat{P}_3$. in %
D=1100, burn-in period=200, c =100, $\zeta^{(0)}$ =maximum likelihood estimates			
Random Reg Imp	0.55	22.39	1.91
NN, Q=10	0.45	26.78	2.17
HDI, Q=10	0.44	26.35	2.35
D=100, burn-in period=10, c =10, $\zeta^{(0)}$ =maximum likelihood estimates			
Random Reg Imp	0.54	22.39	1.91
NN, Q=10	0.45	26.70	2.16
HDI, Q=10	0.46	26.42	2.39
D=1100, burn-in period=200, c =100, $\zeta^{(0)}$ =chosen estimates			
Random Reg Imp	0.53	22.21	1.91
NN, Q=10	0.45	26.79	2.20
HDI, Q=10	0.44	26.38	2.39
D=100, burn-in period=10, c =10, $\zeta^{(0)}$ =chosen estimates			
Random Reg Imp	0.53	22.37	1.91
NN, Q=10	0.46	26.79	2.20
HDI, Q=10	0.44	26.37	2.39

Table 6.42: Estimated percentages of P_1, P_2 and P_3 for 18+ unweighted, using data augmentation based on the assumption of CME using random regression, nearest neighbour and hot deck imputation within classes. The initial starting values for the unknown parameters are the maximum likelihood estimates based on observed data and chosen values. The iterations used are D=1100, c=100 and D=100, c=10.

DA-MAR	$\hat{P}_1$. in %	$\hat{P}_2$. in %	$\hat{P}_3$. in %
D=1100, burn-in period=200, c=100, $\zeta^{(0)}$ =maximum likelihood estimates			
Random Reg Imp	1.22	24.90	1.93
NN	0.50	28.78	2.29
HDI	0.51	26.99	2.49

Table 6.43: Estimated percentages of P_1, P_2 and P_3 for 18+ unweighted, using data augmentation based on the MAR assumption using random regression, nearest neighbour and hot deck imputation within classes. The initial starting values for the unknown parameters are the maximum likelihood estimates based on observed data. The iterations used are D=1100 and c=100.

MAR-based Imputation without Data Augmentation	$\hat{P}_1$. in %	$\hat{P}_2$. in %	$\hat{P}_3$. in %
Random Reg Imp10	1.24	26.27	1.89
NN10	0.50	29.04	2.27
HDIwor10	0.49	29.07	2.33
HDIwr10	0.50	29.11	2.43

Table 6.44: Estimated percentages of P_1, P_2 and P_3 for 18+ unweighted, using random regression, nearest neighbour and hot deck imputation within classes based on the MAR assumption. The initial starting values for the unknown parameters are the maximum likelihood estimates based on observed data. The iterations used are D=1100 and c=100.

Table 6.43 gives the results for data augmentation under the assumption of MAR. As expected for all three imputation methods the results with data augmentation under MAR are very similar to the results for other MAR-based imputation methods not using data augmentation given in table 6.44. For both cases we observe again that the estimates under random regression imputation differ from the estimates under both hot deck imputation methods. The estimate for P_1 under

regression imputation is, with around 1.22%, much higher than for hot deck imputation with around 0.5%. Whereas the estimates for P_2 and P_3 are with approximately 26% and 1.9% lower than for hot deck imputation with approximately 29% and 2.3%.

Regression Imputation, MAR	$\hat{P}_1$. in %	$\hat{P}_2$. in %	$\hat{P}_3$. in %
Random Reg Imp10 (sd=0.3)	2.49	26.79	1.92
Random Reg Imp10 (sd=0.2)	1.37	26.71	1.92
Random Reg Imp10 (sd=0.18 as estimated for LFS data)	1.24	26.27	1.89
Random Reg Imp10 (sd=0.1)	0.75	25.66	1.90
Random reg imp within classes (the variance of the residuals is defined within classes)	0.88	27.31	1.90
Imputing the predicted values (M=1)	0.52	25.25	1.89

Table 6.45: Estimated percentages of P_1, P_2 and P_3 for 18+ unweighted, using random regression imputation where the residuals are drawn from a normal distribution with mean zero and different choices for the standard deviation sd, random regression imputation within classes and regression imputation imputing the predicted values, based on the MAR assumption.

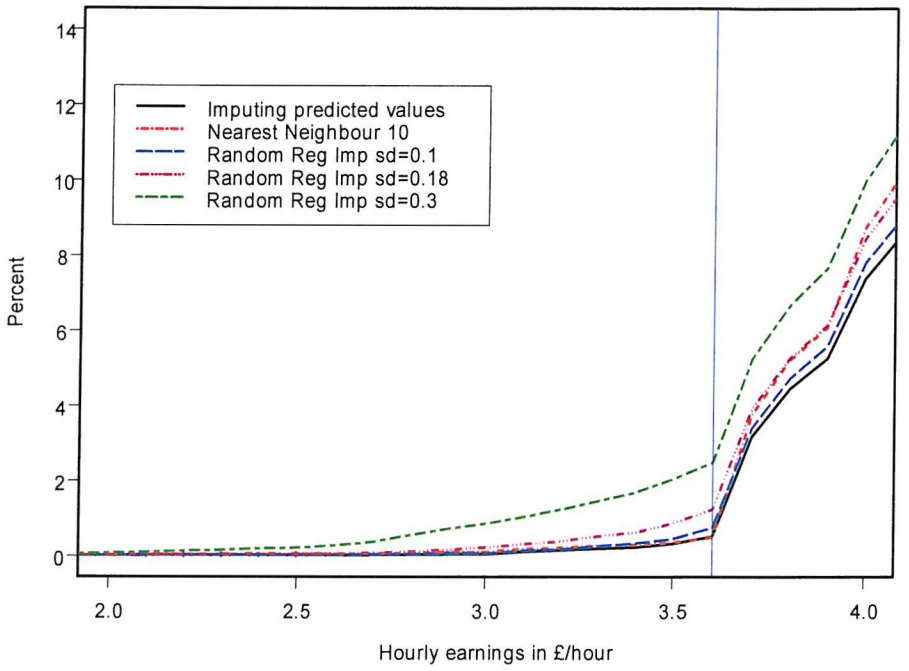

Figure 6.11: Estimated cumulative distribution of hourly earnings from £2 to £4 for 18+ age group, unweighted, under nearest neighbour imputation (NN10), regression imputation imputing the predicted values, random regression imputation adding on residuals drawn from a normal distribution with a standard deviation of 0.1, 0.18 and 0.3, for the March-May 2000 quarter. The standard deviation of 0.18 refers to the observed standard deviation of the residuals from the imputation model. (The methods are all MAR-based).

Since random regression imputation does not seem to give adequate estimates different forms of random regression imputation are analysed. The added on residuals are drawn from a normal distribution with mean zero and different choices for the standard deviation. The results are presented in table 6.45. We can see that with an increasing value for the standard deviation of the residuals, the estimates for P_1 and P_2 increase. The estimate for P_3 stays approximately the same. This illustrates the sensitivity of the imputation method to the choice of the estimated standard deviation of the residuals. Since in section 6.3.5.3 an indication for a non-constant variance of the

residuals of the imputation model is found, random regression imputation within classes is implemented, where the variance of the residuals is defined within classes, i.e. $\hat{\varepsilon}_j \sim N(0,\hat{\sigma}^2_{B_k})$, $j \in \bar{r}$, where $\hat{\sigma}^2_{B_k}$ depends on the class B_k, $k = 1,...,K$. That means constant variance is assumed within classes. The classes are defined based on the range of the predicted values. This method leads to an approximation of the estimates to the estimates obtained under hot deck imputation. However, the estimate for P_1 is still higher than for hot deck imputation. Imputing only the predicted values leads to an underestimation for point estimators $\hat{P}_2$. and $\hat{P}_3$. in comparison to the hot deck methods and is regarded not adequate when estimating distributional quantities. The effect on the estimated distribution functions based on random regression imputation under different choices of standard deviations for the residuals in comparison to nearest neighbour imputation and regression imputation is illustrated in figure 6.11. Imputing the predicted values leads to underestimation of the distribution in comparison to nearest neighbour imputation, whereas with an increasing value of the standard deviation of the residuals random regression imputation leads to overestimation. Due to the dependency of underlying assumptions we conclude that parametric random regression imputation with or without data augmentation is not recommended for practical use.

6.3.5.6 Conclusions on Data Augmentation Method

In this section data augmentation under the common measurement error assumption is derived. It reflects an imputation method under nonignorable nonresponse. Based on the findings of the simulation study we conclude that data augmentation under CME gives approximately unbiased results for the point estimators of interest under ideal model conditions. In addition to the commonly used parametric random regression imputation in the *I*-step of data augmentation two forms of hot deck imputation, hot deck imputation within classes and nearest neighbour imputation, are proposed. This approach has the advantage that certain features of the underlying hourly earnings distributions, such as rounding and truncation effects, can be preserved and that actually observed values are used for imputation. It therefore can overcome some of the problems encountered with random regression imputation in that this approach does not reflect the distribution of hourly earnings adequately. Data augmentation under the CME assumptions performs well for both forms of hot deck imputation under ideal model conditions. Some

evidence is found that the results of hot deck imputation within classes may depend on the choice of classes. It is therefore important to define the classes adequately particularly at the bottom of the distribution of the predicted Y values. Under nearest neighbour imputation one point estimator is found to have a small significant bias, which is thought to be related to the weighted bootstrap. This is in contrast to earlier findings for nearest neighbour imputation under the MAR assumption, where all estimators are approximately unbiased. Data augmentation under CME seems reasonably robust to misspecification of the imputation model. This might be related to the use of the nonresponse model for rejecting or accepting imputed values. As long as the nonresponse model is correctly specified the imputation method is expected to work well. However, for the two hot deck methods a slightly higher increase in the bias is found under misspecification of the imputation model compared to regression imputation which is thought to be related to the use of the weighted bootstrap instead of rejection sampling. Although the two hot deck methods preserve the shape of the distribution better than random regression imputation they may suffer slightly from the use of the weighted bootstrap, which is only an approximation to the rejection sampling procedure. It is found that data augmentation based on CME using hot deck imputation performs well under misspecification of the nonresponse mechanism, such as MAR nonresponse and nonresponse that depends on Y, X and other covariates. For data augmentation using random regression imputation the method seems to be more sensitive to misspecification of the nonresponse mechanism which may be related to the use of rejection sampling instead of the weighted bootstrap method.

Comparing efficiency for DA-CME methods, random regression imputation seems to be more efficient than the two hot deck methods, with nearest neighbour imputation being more efficient than hot deck imputation within classes under ideal model conditions. The results from the simulation study showed that the MI variance estimator performs reasonably well under all three imputation methods taking into account bias and coverage rate. The naive variance estimator underestimates the simulation variance significantly, as expected.

In addition, the method of data augmentation under the CME assumption is applied to the LFS. As expected, the estimates obtained under the CME assumption are found to be slightly lower than under the assumption of MAR. We conclude that overall the imputation method based on data augmentation under the CME assumption performs well. The use of hot deck imputation within the *I*-step showed adequate properties and is recommended for practical use. Although

random regression imputation performs well in the simulation study, in the practical application considered here the method does not seem to produce adequate estimates. The method relies on distributional assumptions, such as the assumption of constant variance, and does not preserve certain features of the distribution well. Therefore random regression imputation is not recommended for practical use. Overall nearest neighbour imputation is recommended either under the assumption of CME or under MAR.

6.4 Conclusion

We investigated several measurement error models, additive and multiplicative, and found that for the LFS application considered here many assumptions made in the classical measurement error model are violated, such as normality of the error terms and constant variance. A common measurement error assumption was proposed as an alternative to the MAR assumption allowing for the fact that the response indicator I depends on true hourly earnings Y, reflecting nonignorable nonresponse. Based on this assumption several estimation methods were discussed such as deconvolution, parametric methods making assumptions of the error distribution and adjustment methods to misclassification. However, these methods have certain disadvantages and do not appear to be appropriate to estimate the distribution of interest. A weighted bootstrap approach was proposed in a Bayesian framework. However, this method also relies on distributional assumptions. It was found that the method did not perform well when applied to LFS data to obtain estimates of low pay. An alternative imputation method using data augmentation under nonignorable nonresponse was proposed, which performed well under certain conditions in the simulation study and in the application to the LFS. It is recommended for practical use.

Chapter 7

Conclusion and Further Work

The aim of this chapter is to summarise the main conclusions of the work undertaken in this research and to highlight the main contributions of this thesis. Possible areas of further research extending the methods presented here are discussed. The investigation of low pay is of interest to a wide range of analysts in economics and the social sciences, in particular because of the introduction of the National Minimum Wage in Great Britain. It is therefore important to provide valid estimation techniques of earnings distributions. The aim of this thesis is to develop and to evaluate different methods for estimating distributions in the presence of measurement error and missing data with particular reference to pay distributions. Some of the methods presented here have already been used by the Office for National Statistics (ONS) to improve hourly earnings data and to obtain estimates of low pay based on large household survey data.

Different methods for correcting for measurement error in a fully observed earnings variable are considered by taking into account information on accurately measured hourly earnings observed on a non-random subsample. To compensate for nonresponse in the correct earnings variable and to effectively correct for the measurement error in the erroneously observed pay variable several imputation methods are proposed. The use of imputation in the presence of nonresponse in surveys, however, requires an identifying assumption of the nonresponse mechanism. Two main assumptions are distinguished in this thesis. The assumption that the data are missing at random

(MAR) is commonly used in the literature, where the nonresponse does not depend on the variable being imputed. Several imputation and weighting methods under this assumption are presented and investigated. However, there are grounds to believe that the assumption of MAR may be unrealistic in the application considered and that in fact the nonresponse depends on the variable that is missing. An alternative assumption of common measurement error (CME) is proposed reflecting nonignorable nonresponse, which may be a more attractive identifying assumption. Since the measurement error here does not follow classical measurement error assumptions the derivation of appropriate imputation methods taking into account the measurement error structure is less straight forward.

One major part of this thesis is the development and investigation of imputation methods initially under the assumption of MAR. The first goal of this thesis is to evaluate hot deck imputation within classes theoretically and empirically. Under certain conditions this imputation method provides approximately unbiased estimates of the point estimators of interest. Since standard variance formulae lead to underestimation in the presence of imputation an important part of this work is the derivation of a valid variance formula allowing for uncertainty due to imputation under this imputation method. A formula for the estimation of the variance is derived. It is shown that this estimator is approximately unbiased under certain conditions. In addition, variance estimation using Rubin's multiple imputation formula is considered. Because the imputation method does not constitute 'proper' multiple imputation, this approach underestimates the variance. An adjusted multiple imputation formula is derived and shown to provide approximately unbiased variance estimation. Some evidence is found, however, that results based on hot deck imputation within classes may be dependent on the choice of classes. This method therefore requires careful consideration of the choice of classes. A wider class of different forms of predictive mean matching imputation as well as propensity score weighting valid under the MAR assumption are therefore compared theoretically and empirically. Stochastic and deterministic forms of imputation methods are investigated with no obvious preference for either form of imputation. In comparison to other imputation and weighting methods investigated here nearest neighbour imputation based on the predicted values of the imputation model shows good properties of the point estimator in terms of bias and efficiency and a good performance under model misspecification. Using repeated imputation in comparison to single imputation leads to a considerable gain in efficiency of the estimator. Comparing nearest neighbour imputation with propensity score weighting slight advantages in terms of the performance under misspecification

and a higher efficiency is found in case of the imputation method. Nearest neighbour imputation based on repeated imputation is therefore recommended for practical use. Applying nearest neighbour imputation and propensity score weighting to LFS data it is found that under comparable conditions the two methods produce very similar estimates of low pay. Both hot deck imputation within classes and nearest neighbour imputation have been implemented by ONS to produce estimates of low pay.

A substantial part of the research described is in the analysis of estimation and imputation techniques under nonignorable nonresponse based on the common measurement error assumption. Methods such as deconvolution and misclassification are briefly reviewed. However, these methods have certain drawbacks and do not seem to provide adequate estimation techniques for the application considered here. An imputation method in a Bayesian framework using a weighted bootstrap approach is proposed. However, it is found that the method strongly relies on distributional assumptions. In particular, the method does not perform well when applied to the Labour Force Survey. An alternative imputation method using data augmentation under the common measurement error assumption is proposed, which is a novel technique to use in this context. The method gives approximately unbiased results under correct model specifications. In addition, the use of hot deck imputation in the imputation step of data augmentation is proposed, which is an extension to the commonly used parametric regression imputation. Data augmentation under the CME assumption using hot deck imputation is found to be reasonably robust against misspecification of the nonresponse mechanism. This is an advantage in comparison to MAR based imputation methods that show a higher increase in the bias under non-MAR nonresponse. However, data augmentation under CME based on hot deck imputation relies on the use of the weighted bootstrap which might lead to an increase in the bias if the imputation model is not correctly specified. Another advantage of using imputation based on data augmentation is the fact that variance estimation is possible using the multiple imputation variance estimation formula. The previously considered MAR based methods which are not proper multiple imputations require the derivation of an expression for the variance.

In the application to the Labour Force Survey data augmentation under the CME assumption leads to estimates of the proportion of low paid that are slightly below the estimates obtained under MAR based methods. Random regression imputation either with or without data augmentation is not recommended for imputation of earnings data since it does not preserve

rounding and truncation effects well. Overall, nearest neighbour imputation either under the MAR assumption or under the CME assumption using augmentation is recommended for practical use.

Possible areas of further research include the following topics. An extension of the work on variance estimation for hot deck imputation within classes, which assumes independence of sample jobs, is variance estimation taking into account the complex design of the UK Labour Force Survey such as stratification and clustering of jobs within households. Since under the assumption of MAR nearest neighbour imputation is found to be a promising imputation method with desirable properties of the point estimator it is of interest to analyse the efficiency of the point estimator under this method as well as variance estimation. A comparison to the efficiency of the point estimator under propensity score weighting as well as variance estimation under the weighting approach are additional extensions of the work.

Another area of research is the further investigation of estimation methods under the common measurement error assumption. One possibility is to explore adjustment methods developed for misclassified categorical data. The use of the misclassification matrix to adjust for misclassification in the proportion of interest based on the derived variable has been discussed in chapter 6. This method can be explored further and the problem of small sample sizes within the categories of interest needs to be addressed. Other methods that make use of the common measurement error assumption may focus on parameter estimation of regression models where the nonresponse depends on the variable subject to missing data. One approach has been proposed by Greenlees et al. (1982) estimating the parameters of the linear regression model predicting Y and the parameters of a nonresponse model dependent on Y analytically using maximum likelihood estimation. Based on the predictions of the two models under the estimated parameters imputed values are obtained using an acceptance-rejection algorithm. It is of interest to apply this approach of maximum likelihood estimation under the common measurement error assumption and to compare the performance of this imputation method with the imputation method derived in chapter 6 using data augmentation.

A very interesting extension of the work is to focus on parameter estimation for the nonresponse model under the common measurement error assumption. If the nonresponse model depending on the variable Y and other covariates W, $f(I \mid Y, W)$, could be estimated correctly, propensity score weighting could be performed where the weights of the point estimator are the inverse of

the estimated probabilities of response from this model. To find the estimates of the parameters of the nonresponse model under nonignorable nonresponse using the common measurement error assumption two approaches are possible. One possibility is to explore methods used in the measurement error literature for estimating coefficients of a regression model where one or more of the covariates are subject to measurement error. The parameters of the model $f(I \mid Y,W)$ cannot be estimated directly by fitting I to (Y,W) since Y is not fully observed. The goal of measurement error modelling is to obtain nearly unbiased estimates of these parameters indirectly by fitting the model $f(I \mid X,W)$, where X is fully observed but measures Y with error, and adjusting the estimated parameters for the effect of measurement error. Substituting X for Y in the nonresponse model without adjustment leads to biased estimates of the coefficients. Possible ways need to be explored how to carry out this adjustment by taking into account the common measurement error assumption and avoiding making further assumptions about the distribution of the measurement error. Methods of measurement error modelling are discussed in Carroll et al. (1995) and Fuller (1987). Another approach of estimating the coefficients of the model $f(I \mid Y,W)$ to carry out propensity score weighting is to impute the missing values of Y and to fit the model directly. The data augmentation procedure proposed in chapter 6 allows for the imputation for Y under the common measurement error assumption. Regarding data augmentation as a form of parameter simulation the data augmentation algorithm derived in chapter 6 generates estimates of the coefficients of the nonresponse model, $f(I=0 \mid Y,W,\gamma)$, from the posterior distribution. Based on the estimated coefficients $\hat{\gamma}$ for this model we can obtain the probabilities of response $\hat{p}_i$, where $\hat{p}_i = 1-\hat{\rho}_i = 1-\hat{f}(I_i=0 \mid y_i,\mathrm{w}_i,\hat{\gamma})$. Using these probabilities to define weights for the respondents it is possible to perform propensity score weighting and to obtain estimates of the proportion of low paid as described in chapter 5. A comparison of propensity score weighting under the CME assumption with propensity score weighting under the MAR assumption is of interest.

Appendices

Appendix: Chapter 2

A 2.1) Estimated Linear Regression Model to Predict the Direct Variable, based on Respondents only (Imputation Model), based on March-May 2000 quarter.

$$\ln(y_i) = \eta_i \hat{\beta} + \hat{\varepsilon}_i ,$$

where η_i is a row-vector of functions of the derived variable and other covariates.

Table of Coefficients:

	Value	Std. Error	t value	Pr(>|t|)
(Intercept)	0.9205	0.1065	8.6452	0.0000
LHE	0.1714	0.0626	2.7367	0.0062
LHEsq	0.1174	0.0052	22.5106	0.0000
SOCcat1	0.4192	0.0693	6.0501	0.0000
SOCcat2	0.3601	0.0578	6.2352	0.0000
SOCcat3	0.1888	0.0517	3.6529	0.0003
SOCcat4	0.2984	0.0553	5.3917	0.0000
SOCcat5	0.1133	0.0494	2.2945	0.0218
SOCcat6	0.3010	0.0538	5.5965	0.0000
SOCcat7	0.2399	0.0529	4.5322	0.0000
SOCcat8	0.2698	0.0518	5.2068	0.0000
PT	0.1115	0.0214	5.2060	0.0000
LTWK	0.2620	0.1497	1.7504	0.0801
Qcat1	-0.0326	0.0126	-2.5965	0.0094
Qcat2	-0.0944	0.0115	-8.2395	0.0000
Qcat3	-0.1112	0.0114	-9.7277	0.0000
Qcat4	-0.1388	0.0120	-11.5407	0.0000
Qcat5	-0.1474	0.0124	-11.8713	0.0000
AGE	0.0068	0.0013	5.3707	0.0000
AGEsq	-0.0001	0.0000	-5.5149	0.0000
EMPMONcat1	0.0000	0.0076	0.0032	0.9975
EMPMONcat2	0.0219	0.0067	3.2595	0.0011
EMPMONcat3	0.0574	0.0063	9.0462	0.0000
HOH	0.0181	0.0065	2.7752	0.0055
MARRIED	0.0230	0.0055	4.1870	0.0000
SIZE	0.1247	0.0195	6.3879	0.0000
INDcat1	0.4539	0.1444	3.1439	0.0017
INDcat2	-0.1922	0.0941	-2.0433	0.0411
INDcat3	0.0663	0.1028	0.6456	0.5186
INDcat4	-0.0664	0.0939	-0.7074	0.4794
INDcat5	-0.1553	0.1002	-1.5501	0.1212
INDcat6	0.0118	0.0957	0.1229	0.9022
INDcat7	-0.0926	0.0937	-0.9882	0.3231
INDcat8	-0.2056	0.0989	-2.0793	0.0376
REGcat1	0.0191	0.0122	1.5564	0.1197
REGcat2	0.0219	0.0123	1.7822	0.0748
REGcat3	0.0138	0.0131	1.0543	0.2918

REGcat4	0.0100	0.0125	0.8047	0.4210
REGcat5	0.0489	0.0127	3.8581	0.0001
REGcat6	0.0979	0.0133	7.3798	0.0000
REGcat7	0.0666	0.0118	5.6304	0.0000
REGcat8	0.0310	0.0124	2.4944	0.0126
REGcat9	0.0298	0.0142	2.0947	0.0362
REGcat10	0.0215	0.0122	1.7612	0.0783
REGcat11	-0.0131	0.0175	-0.7448	0.4564
FEMALE	-0.0291	0.0072	-4.0601	0.0000
USGRS	0.0270	0.0053	5.0659	0.0000
LHE:SOCcat1	-0.1008	0.0304	-3.3175	0.0009
LHE:SOCcat2	-0.1604	0.0271	-5.9282	0.0000
LHE:SOCcat3	-0.1251	0.0260	-4.8023	0.0000
LHE:SOCcat4	-0.1769	0.0271	-6.5209	0.0000
LHE:SOCcat5	-0.1215	0.0252	-4.8158	0.0000
LHE:SOCcat6	-0.2499	0.0291	-8.5759	0.0000
LHE:SOCcat7	-0.1957	0.0265	-7.3730	0.0000
LHE:SOCcat8	-0.2548	0.0275	-9.2650	0.0000
LHE:PT	-0.0892	0.0123	-7.2656	0.0000
LHE:SIZE	-0.0529	0.0112	-4.7244	0.0000
LHE:LTWK	-0.2518	0.0648	-3.8834	0.0001
LHE:INDcat1	-0.1615	0.0759	-2.1272	0.0334
LHE:INDcat2	0.1551	0.0565	2.7447	0.0061
LHE:INDcat3	0.0253	0.0600	0.4216	0.6734
LHE:INDcat4	0.0558	0.0568	0.9823	0.3260
LHE:INDcat5	0.1339	0.0594	2.2545	0.0242
LHE:INDcat6	0.0558	0.0572	0.9749	0.3296
LHE:INDcat7	0.1030	0.0564	1.8271	0.0677
LHE:INDcat8	0.1395	0.0595	2.3434	0.0191

Residual standard error: 0.1824 on 6737 degrees of freedom
Multiple R-Squared: 0.7719
F-statistic: 350.8 on 65 and 6737 degrees of freedom, the p-value is 0

Analysis of Variance Table

	Df	Sum of Sq	Mean Sq	F Value	Pr(F)
LHE	1	600.8350	600.8350	18061.27	0.0000000000
LHEsq	1	45.3164	45.3164	1362.22	0.0000000000
SOCcat	8	61.9314	7.7414	232.71	0.0000000000
PT	1	4.8756	4.8756	146.56	0.0000000000
LTWK	1	5.2246	5.2246	157.05	0.0000000000
Qcat	5	6.2562	1.2512	37.61	0.0000000000
AGE	1	1.9483	1.9483	58.57	0.0000000000
AGEsq	1	2.3169	2.3169	69.65	0.0000000000
EMPMONcat	3	3.4543	1.1514	34.61	0.0000000000
HOH	1	1.1173	1.1173	33.59	0.0000000071
MARRIED	1	0.8660	0.8660	26.03	0.0000003448
SIZE	1	2.3672	2.3672	71.16	0.0000000000
INDcat	8	3.4421	0.4303	12.93	0.0000000000
REGcat	11	4.8908	0.4446	13.37	0.0000000000
FEMALE	1	0.5597	0.5597	16.83	0.0000414479
USGRS	1	0.9749	0.9749	29.31	0.0000000639
LHE:SOCcat	8	6.9538	0.8692	26.13	0.0000000000
LHE:PT	1	1.8176	1.8176	54.64	0.0000000000
LHE:SIZE	1	0.4254	0.4254	12.79	0.0003513676
LHE:LTWK	1	0.4512	0.4512	13.56	0.0002325986
LHE:INDcat	8	2.5753	0.3219	9.68	0.0000000000
Residuals	6737	224.1163	0.0333		

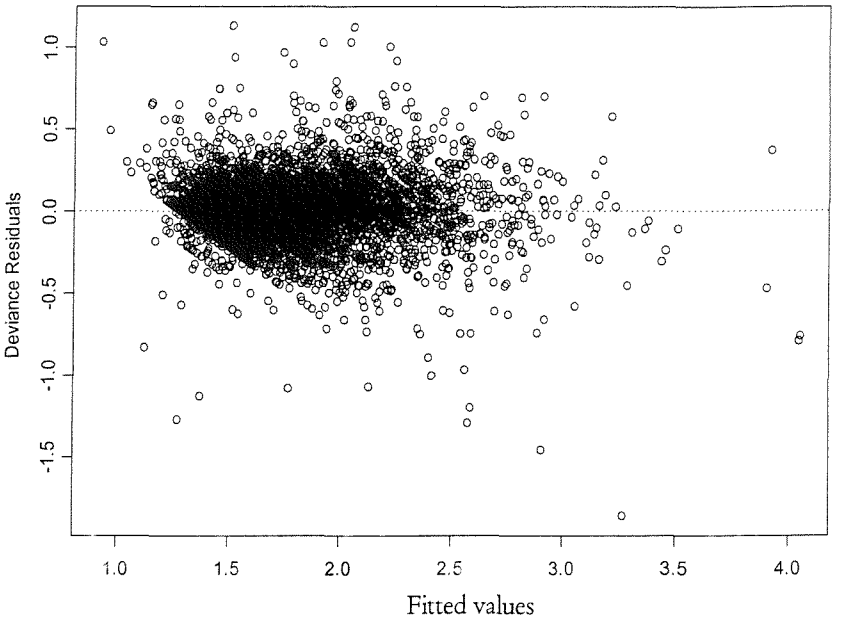

Figure A 2.1.1: Plot of residuals versus fitted values of the imputation model.

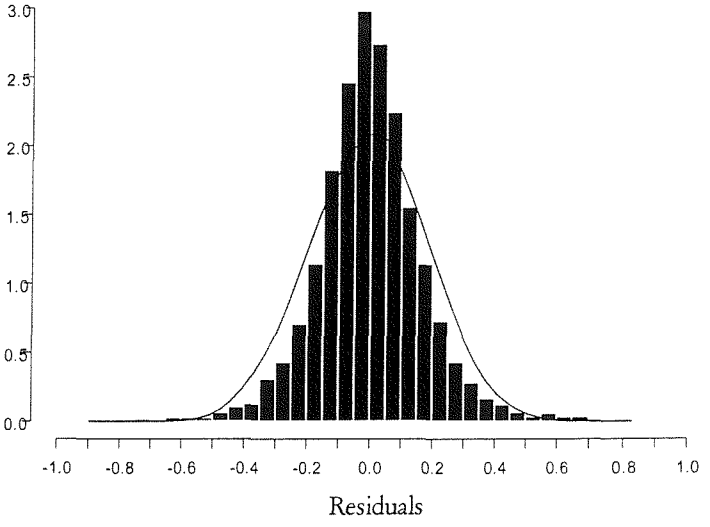

Figure A 2.1.2: Histogram of the residuals of imputation model with normal density curve.

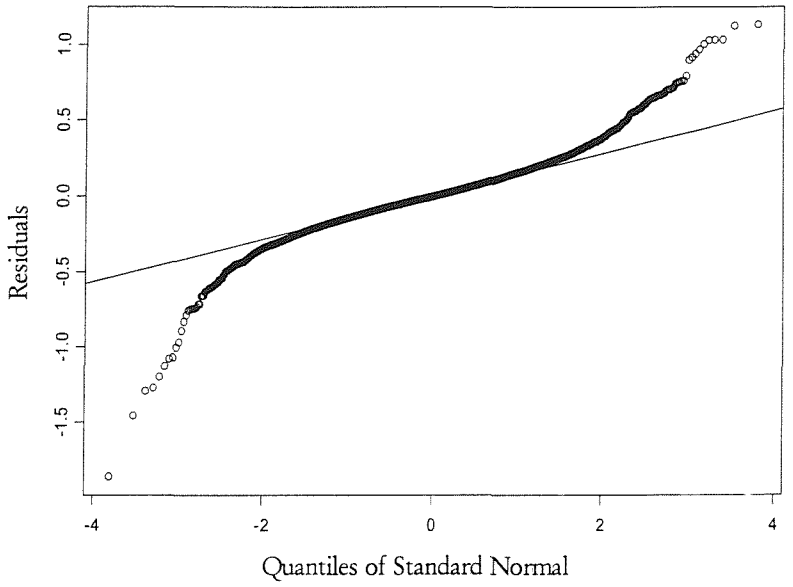

Figure A 2.1.3: Normal probability plot of imputation model.

Appendix: Chapter 3

A 3.1) Estimated Linear Regression Model to Simulate the Derived Variable in the Simulation Study, based on March-May 2000 quarter.

$$\ln(x_i) = \mu_i \hat{\lambda}_r + \hat{\varepsilon}_i,$$

where μ_i is a row-vector of functions of the covariates.

Table of Coefficients:

	Value	Std. Error	t value	Pr(>|t|)
(Intercept)	1.7037	0.0401	42.4737	0.0000
SOC2	-0.0121	0.0147	-0.8187	0.4129
SOC3	-0.1342	0.0150	-8.9169	0.0000
SOC4	-0.3400	0.0141	-24.0849	0.0000
SOC5	-0.3914	0.0151	-25.7704	0.0000
SOC6	-0.4405	0.0162	-27.1237	0.0000
SOC7	-0.3323	0.0203	-16.3224	0.0000
SOC8	-0.4890	0.0156	-31.2346	0.0000
SOC9	-0.5757	0.0206	-27.9394	0.0000
AGE	0.0313	0.0020	15.3577	0.0000
AGEsquared	-0.0003	0.0000	-14.9813	0.0000
Q2	-0.1497	0.0137	-10.8762	0.0000
Q3	-0.2303	0.0123	-18.5960	0.0000
Q4	-0.2850	0.0126	-22.4879	0.0000
Q5	-0.3392	0.0144	-23.5320	0.0000
Q6	-0.4153	0.0153	-27.0104	0.0000
PT	-0.1522	0.0351	-4.3268	0.0000
EMPMON2	0.0133	0.0124	1.0707	0.2842
EMPMON3	0.0511	0.0107	4.7383	0.0000
EMPMON4	0.1510	0.0099	15.1151	0.0000
LTWK	0.7006	0.0679	10.3114	0.0000
HOH	0.0876	0.0101	8.6254	0.0000
MARRIED	0.0643	0.0081	7.8688	0.0000
SIZE	0.1376	0.0074	18.3801	0.0000
FEMALE	-0.1173	0.0104	-11.1950	0.0000
PT:SOC2	0.1656	0.0455	3.6393	0.0002
PT:SOC3	0.0950	0.0427	2.2231	0.0262
PT:SOC4	0.1057	0.0390	2.7080	0.0067
PT:SOC5	-0.0879	0.0629	-1.3972	0.1623
PT:SOC6	0.0001	0.0399	0.0045	0.9963
PT:SOC7	-0.1430	0.0420	-3.4014	0.0006
PT:SOC8	-0.0420	0.0519	-0.8098	0.4180
PT:SOC9	0.0708	0.0427	1.6567	0.0975

```
Residual standard error: 0.4211 on 15689 degrees of freedom
Multiple R-Squared: 0.4595
F-statistic: 416.8 on 32 and 15689 degrees of freedom, the p-value is 0
```

Analysis of Variance Table

	Df	Sum of Sq	Mean Sq	F Value	Pr(F)
SOC2	1	540.8842	540.8842	3049.24011	0.0000
SOC3	1	179.5458	179.5458	1012.1915	0.0000
SOC4	1	1.6733	1.6733	9.4332	0.0021
SOC5	1	3.2018	3.2018	18.0506	0.0000
SOC6	1	90.3275	90.3275	509.2224	0.0000
SOC7	1	149.4651	149.4651	842.6115	0.0000
SOC8	1	107.3562	107.3562	605.2217	0.0000
SOC9	1	561.7027	561.7027	3166.6046	0.0000
AGE	1	32.0831	32.0831	180.8689	0.0000
AGEsquared	1	104.5148	104.5148	589.2034	0.0000
Q2	1	5.1167	5.1167	28.8457	0.0000
Q3	1	11.8355	11.8355	66.7229	0.0000
Q4	1	3.1596	3.1596	17.8125	0.0002
Q5	1	16.9264	16.9264	95.4230	0.0000
Q6	1	171.0494	171.0494	964.2929	0.0000
PT	1	113.8462	113.8462	641.8094	0.0000
EMPMON2	1	11.0804	11.0804	62.4660	0.0000
EMPMON3	1	7.4928	7.4928	42.2410	0.0000
EMPMON4	1	48.3013	48.3013	272.2991	0.0000
LTWK	1	18.0503	18.0503	101.7588	0.0000
HOH	1	63.5763	63.5763	358.4121	0.0000
MARRIED	1	22.6063	22.6063	127.4434	0.0000
SIZE	1	59.1101	59.1101	333.2340	0.0000
FEMALE	1	23.1341	23.1341	130.4187	0.0000
PT:SOC2	1	3.8350	3.8350	21.6201	0.0000
PT:SOC3	1	1.6753	1.6753	9.4448	0.0021
PT:SOC4	1	7.0093	7.0093	39.5153	0.0000
PT:SOC5	1	0.2656	0.2656	1.4974	0.2210
PT:SOC6	1	0.3729	0.3729	2.1026	0.1470
PT:SOC7	1	5.9253	5.9253	33.4040	0.0000
PT:SOC8	1	0.7492	0.7492	4.2238	0.0398
PT:SOC9	1	0.4868	0.4868	2.7447	0.0975

A 3.2) Logistic Regression for Modeling the Nonresponse (Model A3) on the Variable Y in the LFS, based on March-May 2000 quarter.

This model includes $\ln(Y)$, $\ln(Y)^2$, other explanatory variables of interest and two-way interactions.

Table of Coefficients:

	Value	Std. Error	t value
(Intercept)	-0.912023443	0.40892949	-2.2302706
LHE	0.398128606	0.20667349	1.9263651
LHEsquared	-0.407888457	0.04049653	-10.0721841
SOCcat1	-1.520777163	0.48535108	-3.1333548
SOCcat2	-0.163540196	0.41510434	-0.3939737
SOCcat3	1.085178227	0.36596454	2.9652552
SOCcat4	0.313548817	0.38803962	0.8080330
SOCcat5	2.140319869	0.36637293	5.8419159
SOCcat6	3.249175581	0.43285915	7.5063114
SOCcat7	1.137717078	0.39854424	2.8546820
SOCcat8	2.116546551	0.41446221	5.1067298
PT	-0.008578994	0.18356208	-0.0467362
LTWK	1.398779931	0.47374241	2.9526171
EMPMONcat1	-0.146508938	0.07255539	-2.0192702
EMPMONcat2	-0.373229528	0.06254793	-5.9670960
EMPMONcat3	-0.421140909	0.05658455	-7.4426835
ADDTBP	0.515840814	0.04699016	10.9776354
PROXYcat1	-0.694320532	0.05127061	-13.5422715
PROXYcat2	-0.817738587	0.09352918	-8.7431386
HOH	-0.015418605	0.05730990	-0.2690391
SIZE	0.111935033	0.04500749	2.4870313
USGRS	-0.689860853	0.05537307	-12.4584174
FEMALE	0.101040249	0.06227572	1.6224660
INDcat1	0.979072019	0.27869593	3.5130474
INDcat2	0.951020826	0.21665389	4.3895857
INDcat3	0.856136860	0.22604281	3.7874988
INDcat4	0.915462881	0.21681198	4.2223815
INDcat5	0.623663430	0.22294239	2.7974197
INDcat6	0.409694947	0.21924622	1.8686523
INDcat7	0.642712144	0.21744522	2.9557428
INDcat8	0.658502811	0.22896986	2.8759366
Qcat1	0.549129663	0.08975892	6.1178286
Qcat2	0.588052213	0.08269284	7.1112832
Qcat3	0.588311058	0.08385367	7.0159249
Qcat4	0.810467547	0.09200073	8.8093598
Qcat5	0.733658916	0.09643406	7.6078816
REGcat1	-0.157724696	0.11055836	-1.4266194
REGcat2	0.129993415	0.11370566	1.1432449
REGcat3	-0.274535239	0.11837965	-2.3191084
REGcat4	-0.243133785	0.11231682	-2.1647139
REGcat5	-0.154898414	0.11417293	-1.3567000
REGcat6	-0.267319324	0.11605591	-2.3033667
REGcat7	0.012608110	0.10725273	0.1175551
REGcat8	0.063401255	0.11445993	0.5539166

REGcat9	-0.159666718	0.12916779	-1.2361187
REGcat10	0.106636154	0.11231717	0.9494199
REGcat11	-0.335404884	0.14985519	-2.2381933
LHE:SOCcat1	0.665705969	0.20635931	3.2259555
LHE:SOCcat2	0.345151938	0.18807720	1.8351610
LHE:SOCcat3	-0.388861711	0.17957188	-2.1654933
LHE:SOCcat4	0.587304273	0.18447673	3.1836225
LHE:SOCcat5	-0.584341839	0.18287846	-3.1952469
LHE:SOCcat6	-1.352313160	0.22771507	-5.9386194
LHE:SOCcat7	0.238229656	0.19677980	1.2106408
LHE:SOCcat8	-0.384992477	0.22043780	-1.7464903
LHE:PT	0.667023875	0.09684565	6.8874944

Analysis of Deviance Table:

	Df	Deviance	Resid. Df	Resid. Dev
NULL			15721	21509.66
LHE	1	2624.604	15720	18885.06
LHEsquared	1	258.618	15719	18626.44
SOCcat	8	1084.181	15711	17542.26
PT	1	776.070	15710	16766.19
LTWK	1	9.778	15709	16756.41
EMPMONcat	3	37.420	15706	16718.99
ADDTBP	1	297.814	15705	16421.18
PROXYcat	2	256.806	15703	16164.37
HOH	1	5.942	15702	16158.43
SIZE	1	7.836	15701	16150.60
USGRS	1	161.078	15700	15989.52
FEMALE	1	4.963	15699	15984.55
INDcat	8	101.679	15691	15882.88
Qcat	5	80.460	15686	15802.42
REGcat	11	59.983	15675	15742.43
LHE:SOCcat	8	112.149	15667	15630.28
LHE:PT	1	46.089	15666	15584.19

Appendix: Chapter 4

A 4.1) Proof of Result 4.1:

To prove result 4.1 we need to derive three components of the total variance referring to imputation, response and sampling variability.

$$\mathrm{var}_{(1)DRI}(\hat{\bar{P}}_{\cdot}) = \mathrm{var}_{(1)DRI}\left[\frac{1}{M}\sum_{m=1}^{M}\hat{P}_{\cdot m}\right]$$

$$= E_{DR}\mathrm{var}_{I}\left[\frac{1}{M}\sum_{m=1}^{M}\hat{P}_{\cdot m}\right] + \mathrm{var}_{D}E_{RI}\left[\frac{1}{M}\sum_{m=1}^{M}\hat{P}_{\cdot m}\right] + E_{D}\mathrm{var}_{R}E_{I}\left[\frac{1}{M}\sum_{m=1}^{M}\hat{P}_{\cdot m}\right]$$

1.) 2.) 3.)

1.)

$$\mathrm{var}_{I}\left[\frac{1}{M}\sum_{m=1}^{M}\hat{P}_{\cdot m}\right] = \mathrm{var}_{I}\left[\frac{1}{M}\sum_{m=1}^{M}\frac{1}{n}\left[\sum_{i\in r}z_{i} + \sum_{i\in\bar{r}}\hat{z}_{mi}\right]\right] = \frac{1}{M^{2}n^{2}}\sum_{m=1}^{M}\sum_{i\in\bar{r}}\mathrm{var}_{I}(\hat{z}_{mi})$$

since $\hat{z}_{mi}$ independent for all m and i because of selection of donors with replacement

$$= \frac{1}{M^{2}n^{2}}\sum_{m=1}^{M}\sum_{i\in\bar{r}}E_{I}(\hat{z}_{mi})(1-E_{I}(\hat{z}_{mi})) = \frac{1}{M^{2}n^{2}}\sum_{m=1}^{M}\sum_{i\in\bar{r}}\bar{z}_{rk(i)}(1-\bar{z}_{rk(i)})$$

$$= \frac{1}{Mn^{2}}\sum_{k}\sum_{i\in B_{k}}\bar{z}_{rk}(1-\bar{z}_{rk})I(i\in\bar{r}) = \frac{1}{Mn^{2}}\sum_{k}\bar{z}_{rk}(1-\bar{z}_{rk})n_{\bar{r}k} = \frac{1}{Mn^{2}}\sum_{k}\bar{z}_{rk}(1-\bar{z}_{rk})(n_{k}-n_{rk})$$

$$E_{DR}\mathrm{var}_{I}\left[\frac{1}{M}\sum_{m=1}^{M}\hat{P}_{\cdot m}\right] = E_{DR}\left[\frac{1}{Mn^{2}}\sum_{k}\bar{z}_{rk}(1-\bar{z}_{rk})(n_{k}-n_{rk})\right]$$

We regard E_R as being made up of two components, such that $E_R = E_{R_2R_1}$ where R_1 generates n_{rk}, the number of people responding in class B_k. The term n_{rk} is binomially distributed with parameters (n_k, π_k). It follows $E_{R_1}(n_{rk}) = n_k\pi_k$. R_2 is a SRS of size n_{rk} from n_k. It follows using Taylor approximation:

$$\doteq \frac{1}{Mn^2}\sum_k E_{DR}(\bar{z}_{rk}(1-\bar{z}_{rk}))\, E_{DR_1}(n_k - n_{rk})$$

$$= \frac{1}{Mn^2}\sum_k E_{DR}(\bar{z}_{rk}(1-\bar{z}_{rk}))\, E_D(n_k)(1-\pi_k)$$

$$= \frac{1}{MnN}\sum_k N_k(1-\pi_k)\left(E_{DR}(\bar{z}_{rk}) - E_{DR}(\bar{z}_{rk})^2 - \mathrm{var}_{DR}(\bar{z}_{rk})\right)$$

Note:

$$E_{DR}(\bar{z}_{rk}) \doteq E_D(\bar{z}_{sk}) = E_D\left[\frac{\sum_{i\in B_{Uk}} z_i I(i\in s)}{\sum_{i\in B_{Uk}} I(i\in s)}\right] \doteq \frac{1}{N_k}\sum_{i\in B_{Uk}} z_i = \bar{z}_{Uk}$$

(8.1)

since $E_R(\bar{z}_{rk}) \doteq \bar{z}_{sk}$ (see proof of result 3.1, equation (3.9)) and using Taylor approximations as in (3.14).

$$E_D(n_k) = E_D\left(\sum_{i\in B_{Uk}} I(i\in s)\right) = \frac{N_k n}{N}\ .$$

$$\mathrm{var}_{DR}(\bar{z}_{rk}) = E_D\mathrm{var}_R(\bar{z}_{rk}) + \mathrm{var}_D E_R(\bar{z}_{rk}) = E_D\mathrm{var}_R(\bar{z}_{rk}) + \mathrm{var}_D(\bar{z}_{sk})$$

$$\text{where } \mathrm{var}_R(\bar{z}_{rk}) = \left(\frac{1}{\pi_k n_k} - \frac{1}{n_k}\right)\bar{z}_{sk}(1-\bar{z}_{sk}) \qquad (8.2)$$

$$\text{since } \mathrm{var}_{BE}(\bar{x}_s) = \left(\frac{1}{\alpha N} - \frac{1}{N}\right)S^2_{xU} \text{ under a Bernoulli sampling}$$

design with $pr(i \in s) = \alpha$ (Särndal et al., 1992, p. 65, 3.27)

$$= E_D\left[\left(\frac{1}{\pi_k n_k} - \frac{1}{n_k}\right)\bar{z}_{sk}(1-\bar{z}_{sk})\right] + \mathrm{var}_D(\bar{z}_{sk})$$

$$= \left(\frac{N}{N_k n \pi_k} - \frac{N}{N_k n}\right)\left(E_D(\bar{z}_{sk}) - E_D(\bar{z}_{sk})^2 - \mathrm{var}_D(\bar{z}_{sk})\right) + \mathrm{var}_D(\bar{z}_{sk})$$

$$\text{Note: } \mathrm{var}_D(\bar{z}_{sk}) = \frac{N}{N_k n}\,\bar{z}_{Uk}(1-\bar{z}_{Uk}) \tag{8.3}$$

$$= \left[\frac{N(1-\pi_k)}{N_k n \pi_k}\left(1 - \frac{N}{N_k n}\right) + \frac{N}{N_k n}\right]\bar{z}_{Uk}(1-\bar{z}_{Uk})$$

$$\doteq \left[\frac{N}{\pi_k N_k n}\right]\bar{z}_{Uk}(1-\bar{z}_{Uk}) \quad \text{since } \frac{N}{N_k n} \text{ small}$$

It follows that:

$$E_{DR}\mathrm{var}_I\left[\frac{1}{M}\sum_{m=1}^{M}\hat{P}_{\cdot m}\right] \doteq \frac{1}{MnN}\sum_k N_k(1-\pi_k)\left(\bar{z}_{Uk} - \bar{z}_{Uk}^{\,2} - \frac{N}{N_k n \pi_k}\bar{z}_{Uk}(1-\bar{z}_{Uk})\right)$$

$$= \frac{1}{MnN}\sum_k N_k \bar{z}_{Uk}(1-\bar{z}_{Uk})(1-\pi_k)\left(1 - \frac{N}{N_k n \pi_k}\right).$$

2.)

$$E_{RI}\left[\frac{1}{M}\sum_{m=1}^{M}\hat{P}_{\cdot m}\right] = \frac{1}{M}\sum_{m=1}^{M}E_{RI}(\hat{P}_{\cdot m}) \doteq \frac{1}{M}\sum_{m=1}^{M}\frac{1}{n}\sum_{i\in U} z_i I(i \in s) = \frac{1}{M}\sum_{m=1}^{M}\bar{z}_s = \bar{z}_s$$

(see proof of result 3.1)

$$\mathrm{var}_D\, E_{RI}\left[\frac{1}{M}\sum_{m=1}^{M}\hat{P}_{\cdot m}\right] = \mathrm{var}_D(\bar{z}_s) = \frac{1}{n}\,\bar{z}_U(1-\bar{z}_U) = \frac{P(1-P)}{n}.$$

3.)

$$E_I\left[\frac{1}{M}\sum_{m=1}^{M}\hat{P}_{\cdot m}\right] = \frac{1}{M}\sum_{m=1}^{M}E_I(\hat{P}_{\cdot m}) \doteq \frac{1}{M}\sum_{m=1}^{M}\frac{1}{n}\sum_k\sum_{i\in B_k}\bar{z}_{rk} = \frac{1}{n}\sum_k n_k\bar{z}_{rk}$$

(see proof of result 3.1)

$$\mathrm{var}_R E_I\left[\frac{1}{M}\sum_{m=1}^{M}\hat{P}_{\cdot m}\right] = \mathrm{var}_R\left[\frac{1}{n}\sum_k n_k\bar{z}_{rk}\right] = \frac{1}{n^2}\sum_k n_k^2\,\mathrm{var}_R(\bar{z}_{rk})$$

$$\text{where } \mathrm{var}_R(\bar{z}_{rk}) = \left(\frac{1}{\pi_k n_k} - \frac{1}{n_k}\right)\bar{z}_{sk}(1-\bar{z}_{sk}) \qquad \text{(see (8.2))}$$

$$= \frac{1}{n^2}\sum_k n_k\left(\frac{1}{\pi_k}-1\right)\bar{z}_{sk}(1-\bar{z}_{sk})$$

$$E_D\mathrm{var}_R E_I\left[\frac{1}{M}\sum_{m=1}^{M}\hat{P}_{\cdot m}\right] \doteq E_D\left[\frac{1}{n^2}\sum_k\sum_{i\in B_{Uk}} I(i\in s)\left(\frac{1}{\pi_k}-1\right)\bar{z}_{sk}(1-\bar{z}_{sk})\right]$$

$$= \frac{1}{nN}\sum_k\sum_{i\in B_{Uk}}\left(\frac{1}{\pi_k}-1\right)\left(E_D(\bar{z}_{sk}) - E_D(\bar{z}_{sk})^2 - \mathrm{var}_D(\bar{z}_{sk})\right)$$

$$= \frac{1}{nN}\sum_k N_k\left(\frac{1}{\pi_k}-1\right)\left(\bar{z}_{Uk} - \bar{z}_{Uk}^{\,2} - \frac{N}{nN_k}\bar{z}_{Uk}(1-\bar{z}_{Uk})\right) \qquad \text{(see (8.3))}$$

$$= \frac{1}{nN}\sum_k N_k\bar{z}_{Uk}(1-\bar{z}_{Uk})\left(\frac{1}{\pi_k}-1\right)\left(1-\frac{N}{nN_k}\right).$$

It follows:

$$\mathrm{var}_{(1)DRI}(\hat{P}_{\cdot}) = \mathrm{var}_{(1)DRI}\left[\frac{1}{M}\sum_{m=1}^{M}\hat{P}_{\cdot m}\right] = (1.) + (2.) + (3.) = V_{imp} + V_{sam} + V_{res}$$

$$\doteq \frac{1}{MnN}\sum_k N_k \bar{z}_{Uk}(1-\bar{z}_{Uk})(1-\pi_k)\left(1-\frac{N}{N_k n\pi_k}\right)$$

$$+\frac{P(1-P)}{n}$$

$$+\frac{1}{nN}\sum_k N_k \bar{z}_{Uk}(1-\bar{z}_{Uk})\left(\frac{1}{\pi_k}-1\right)\left(1-\frac{N}{nN_k}\right)$$

❑

A 4.2) Proof of Result 4.3.

We have to show that $E_{DRI}(\hat{\text{var}}_{(1)DRI}(\hat{P}_\cdot)) \doteq \text{var}_{(1)DRI}(\hat{P}_\cdot)$

We write

$$\text{var}_{(1)DRI}(\hat{P}_\cdot) = V_{sam} + A$$

$$\doteq \frac{P(1-P)}{n} + \frac{1}{nN}\sum_k N_k \bar{z}_{Uk}(1-\bar{z}_{Uk})\left(\frac{1}{M}(1-\pi_k)(1-\frac{N}{N_k n\pi_k})+(\frac{1}{\pi_k}-1)(1-\frac{N}{nN_k})\right),$$

where A is defined as $A = V_{imp} + V_{res}$.

We seek approximately unbiased estimators of the two components V_{sam} and A, such that

$$\hat{\text{var}}_{(1)DRI}(\hat{P}_\cdot) = \hat{V}_{sam} + \hat{A}\ .$$

We define the first term in $\hat{\text{var}}_{(1)DRI}(\hat{P}_\cdot)$ as

$\hat{V}_{sam} = \frac{\hat{P}_\cdot(1-\hat{P}_\cdot)}{n-1}$ and the second term as

$$\hat{A} = \frac{1}{n(n-1)} \sum_k n_k \bar{z}_{\cdot sk}(1-\bar{z}_{\cdot sk}) \left(\frac{1}{M}(1-\hat{\pi}_k)(1-\frac{1}{n_{rk}}) + (\frac{1}{\hat{\pi}_k}-1)(1-\frac{1}{n_k} \right)$$

$$* \left[1 - \frac{1}{n_k} \left(1 + \frac{1}{M}(1-\hat{\pi}_k)(1-\frac{1}{n_{rk}}) + (\frac{1}{\hat{\pi}_k}-1)(1-\frac{1}{n_k}) \right) \right]^{-1},$$

We now need to show that $\hat{V}_{sam}$ and $\hat{A}$ are approximately unbiased estimators of V_{sam} and A respectively.

1.) to show: $E_{DRI}(\hat{A}) \doteq A$

$$E_{DRI}(\hat{A}) = E_{DRI} \left(\frac{1}{n(n-1)} \sum_k n_k \bar{z}_{\cdot sk}(1-\bar{z}_{\cdot sk}) \left(\frac{1}{M}(1-\hat{\pi}_k)(1-\frac{1}{n_{rk}}) + (\frac{1}{\hat{\pi}_k}-1)(1-\frac{1}{n_k} \right) \right.$$

$$\left. * \left[1 - \frac{1}{n_k} \left(1 + \frac{1}{M}(1-\hat{\pi}_k)(1-\frac{1}{n_{rk}}) + (\frac{1}{\hat{\pi}_k}-1)(1-\frac{1}{n_k}) \right) \right]^{-1} \right)$$

$$\doteq \frac{1}{nN} \sum_k N_k \left[E_{DRI}(\bar{z}_{\cdot sk}) - E_{DRI}(\bar{z}_{\cdot sk})^2 - \mathrm{var}_{DRI}(\bar{z}_{\cdot sk}) \right]$$

$$* \left(\frac{1}{M}(1-\pi_k)(1-\frac{N}{N_k n \pi_k}) + (\frac{1}{\pi_k}-1)(1-\frac{N}{N_k n}) \right)$$

$$* \left[1 - \frac{N}{N_k n} \left(1 + \frac{1}{M}(1-\pi_k)(1-\frac{N}{N_k n \pi_k}) + (\frac{1}{\pi_k}-1)(1-\frac{N}{N_k n}) \right) \right]^{-1}$$

Note:

$E_{DRI}(\bar{z}_{\cdot sk}) \doteq \bar{z}_{Uk}$ and

$$\mathrm{var}_{(1)DRI}(\bar{z}_{\cdot sk}) \doteq \frac{N}{N_k n} \bar{z}_{Uk}(1-\bar{z}_{Uk}) \left[1 + \frac{1}{M}(1-\pi_k) \left(1 - \frac{N}{N_k n \pi_k} \right) + (\frac{1}{\pi_k}-1) \left(1 - \frac{N}{N_k n} \right) \right]$$

(these formulae are shown at the end of this proof)

$$= \frac{1}{nN}\sum_k N_k\left(\bar{z}_{Uk} - \bar{z}_{Uk}^{\,2} - \mathrm{var}_{(1)DRI}(\bar{z}_{\cdot sk})\right)$$

$$*\left(\frac{1}{M}(1-\pi_k)(1-\frac{N}{N_k n\pi_k}) + (\frac{1}{\pi_k}-1)(1-\frac{N}{N_k n})\right)$$

$$*\left[1-\frac{N}{N_k n}\left(1+\frac{1}{M}(1-\pi_k)(1-\frac{N}{N_k n\pi_k}) + (\frac{1}{\pi_k}-1)(1-\frac{N}{N_k n})\right)\right]^{-1}$$

$$\doteq \frac{1}{nN}\sum_k N_k \bar{z}_{Uk}(1-\bar{z}_{Uk})\left[1-\frac{N}{N_k n}\left(1+\frac{1}{M}(1-\pi_k)(1-\frac{N}{N_k n\pi_k}) + (\frac{1}{\pi_k}-1)(1-\frac{N}{N_k n})\right)\right]$$

$$*\left(\frac{1}{M}(1-\pi_k)(1-\frac{N}{N_k n\pi_k}) + (\frac{1}{\pi_k}-1)(1-\frac{N}{N_k n})\right)$$

$$*\left[1-\frac{N}{N_k n}\left(1+\frac{1}{M}(1-\pi_k)(1-\frac{N}{N_k n\pi_k}) + (\frac{1}{\pi_k}-1)(1-\frac{N}{N_k n})\right)\right]^{-1}$$

$$= \frac{1}{nN}\sum_k N_k \bar{z}_{Uk}(1-\bar{z}_{Uk})\left(\frac{1}{M}(1-\pi_k)(1-\frac{N}{N_k n\pi_k}) + (\frac{1}{\pi_k}-1)(1-\frac{N}{N_k n})\right)$$

$$\doteq A$$

2.) to show: $E_{DRI}(\hat{V}_{sam}) \doteq V_{sam}$

$$E_{DRI}(\hat{V}_{sam}) = E_{DRI}\left(\frac{\hat{P}_\cdot(1-\hat{P}_\cdot)}{n-1}\right) \doteq \frac{1}{n-1}\left(E_{DRI}(\hat{P}_\cdot) - \left[E_{DRI}(\hat{P}_\cdot)\right]^2 - \mathrm{var}_{(1)DRI}(\hat{P}_\cdot)\right)$$

$$\doteq \frac{1}{n-1}\left(P - P^2 - \left[\frac{P(1-P)}{n} + A\right]\right) = \frac{1}{n-1}\left(P(1-P)(1-\frac{1}{n}) - A\right) = \frac{P(1-P)}{n-1}(\frac{n-1}{n}) - \frac{A}{n-1}$$

$$= \frac{P(1-P)}{n} - \frac{A}{n-1} \doteq \frac{P(1-P)}{n} \doteq V_{sam}, \text{ since } \frac{A}{n-1} \text{ negligible.}$$

Additions:

To show: $E_{DRI}(\bar{z}_{\cdot sk}) \doteq \bar{z}_{Uk}$

Proof

$$E_{DRI}(\bar{z}_{\cdot sk}) = E_{DRI}\left[\frac{1}{M}\sum_{m=1}^{M} z_{\cdot skm}\right] = \frac{1}{M}\sum_{m=1}^{M} E_{DR}\left[\frac{1}{n_k}\left(\sum_{i\in B_k \cap r} z_i + \sum_{i\in B_k \cap \bar{r}} E_I(\hat{z}_{mi})\right)\right]$$

$$= E_{DR}\left[\frac{1}{n_k}\left(\sum_{i\in B_k \cap r} z_i + \sum_{i\in B_k \cap \bar{r}} z_{rk(i)}\right)\right] = E_{DR}\left[\frac{1}{n_k}\sum_{i\in B_k} \bar{z}_{ik}\right] = E_{DR}(\bar{z}_{rk}) \doteq E_D(\bar{z}_{sk}) \doteq \bar{z}_{Uk} \qquad (8.4)$$

(using the same arguments as in proof of result 3.1, equation (3.9) and (8.1))

To show: $\mathrm{var}_{(1)DRI}(\bar{z}_{\cdot sk}) \doteq \frac{N}{N_k n}\bar{z}_{Uk}(1-\bar{z}_{Uk})\left[1+\frac{1}{M}(1-\pi_k)\left(1-\frac{N}{N_k n \pi_k}\right)+\left(\frac{1}{\pi_k}-1\right)\left(1-\frac{N}{N_k n}\right)\right]$

Proof

$$\mathrm{var}_{(1)DRI}(\bar{z}_{\cdot sk}) = E_{DR}\mathrm{var}_I(\bar{z}_{\cdot sk}) + \mathrm{var}_D E_{RI}(\bar{z}_{\cdot sk}) + E_D \mathrm{var}_R E_I(\bar{z}_{\cdot sk})$$

1.) 2.) 3.)

1.)

$$\mathrm{var}_I(\bar{z}_{\cdot sk}) = \frac{1}{M^2}\sum_{m=1}^{M}\mathrm{var}_I(\bar{z}_{\cdot skm}) = \frac{1}{M^2}\sum_{m=1}^{M}\frac{1}{n_k^2}\sum_{i\in B_k\cap\bar{r}}\mathrm{var}_I(\hat{z}_{mi}) = \frac{1}{Mn_k^2}\sum_{i\in B_k\cap\bar{r}}\bar{z}_{rk}(1-\bar{z}_{rk})$$

since $\hat{z}_{mi}$ independent for all i and m because of selection of donors with replacement

$$E_{DR}\mathrm{var}_I(\bar{z}_{\cdot sk}) = E_{DR}\left[\frac{1}{Mn_k^2}\sum_{i\in B_k}\bar{z}_{rk}(1-\bar{z}_{rk})I(i\in\bar{r})\right]$$

$$\doteq \frac{N}{MN_k n}(1-\pi_k)\left[E_{DR}(\bar{z}_{rk}) - E_{DR}(\bar{z}_{rk})^2 - \mathrm{var}_{DR}(\bar{z}_{rk})\right]$$

$$\doteq \frac{N}{MN_k n}\bar{z}_{Uk}(1-\bar{z}_{Uk})(1-\pi_k)\left(1-\frac{N}{N_k n\pi_k}\right)$$

(using the formulae derived in the proof of result 4.1, equation (8.2) and (8.3))

2.)

$E_{RI}(\bar{z}_{\cdot sk}) \doteq \bar{z}_{sk}$ (see (8.4))

$\mathrm{var}_D\, E_{RI}(\bar{z}_{\cdot sk}) \doteq \mathrm{var}_D(\bar{z}_{sk}) = \frac{N}{N_k n}\bar{z}_{Uk}(1-\bar{z}_{Uk})$ (see (8.3))

3.)

$E_I(\bar{z}_{\cdot rk}) \doteq \bar{z}_{rk}$ (see (8.4))

$$E_D \mathrm{var}_R E_I(\bar{z}_{\cdot sk}) \doteq E_D \mathrm{var}_R(\bar{z}_{rk}) \doteq E_D\left[\left(\frac{1}{\pi_k n_k}-\frac{1}{n_k}\right)\bar{z}_{sk}(1-\bar{z}_{sk})\right]$$

$$=\left(\frac{N}{N_k n\pi_k}-\frac{N}{N_k n}\right)\left(E_D(\bar{z}_{sk})-E_D(\bar{z}_{sk})^2-\mathrm{var}_D(\bar{z}_{sk})\right)=\frac{N}{N_k n}\bar{z}_{Uk}(1-\bar{z}_{Uk})\left(\frac{1}{\pi_k}-1\right)\left(1-\frac{N}{N_k n}\right)$$

(using the formulae derived in the proof of result 4.1, equations (8.2) and (8.3))

It follows:

$\mathrm{var}_{(1)DRI}(\bar{z}_{\cdot sk}) = (1.) + (2.) + (3.)$

$$\doteq \frac{N}{MN_k n}\bar{z}_{Uk}(1-\bar{z}_{Uk})(1-\pi_k)\left(1-\frac{N}{N_k n\pi_k}\right)+\frac{N}{N_k n}\bar{z}_{Uk}(1-\bar{z}_{Uk})$$

$$+\frac{N}{N_k n}\bar{z}_{Uk}(1-\bar{z}_{Uk})\left(\frac{1}{\pi_k}-1\right)\left(1-\frac{N}{N_k n}\right)$$

$$= \frac{N}{N_k n}\bar{z}_{Uk}(1-\bar{z}_{Uk})\left[1+\frac{1}{M}(1-\pi_k)\left(1-\frac{N}{N_k n\pi_k}\right)+\left(\frac{1}{\pi_k}-1\right)\left(1-\frac{N}{N_k n}\right)\right] \qquad \square$$

A4.3) Expectation of Multiple Imputation Variance Estimator

$$E_{DRI}(\hat{\text{var}}_{MI}(\hat{P}_{\cdot})) = E_{DRI}\left[\frac{1}{M}\sum_{m=1}^{M}\frac{\hat{P}_{\cdot m}(1-\hat{P}_{\cdot m})}{n-1}\right] + E_{DRI}\left[(1+\frac{1}{M})\frac{1}{M-1}\sum_{m=1}^{M}\left(\hat{P}_{\cdot m}-\hat{P}_{\cdot}\right)^2\right]$$

1.) 2.)

1.)

$$E_{DRI}\left[\frac{1}{M}\sum_{m=1}^{M}\frac{\hat{P}_{\cdot m}(1-\hat{P}_{\cdot m})}{n-1}\right] = \frac{1}{M}\sum_{m=1}^{M}\frac{1}{n-1}\left[E_{DRI}(\hat{P}_{\cdot m}) - E_{DRI}(\hat{P}_{\cdot m})^2 - \text{var}_{(2)DRI}(\hat{P}_{\cdot m})\right]$$

$$\doteq \frac{1}{n-1}\left[P(1-P) - \frac{P(1-P)}{n} - \frac{1}{nN}\sum_{k} N_k \bar{z}_{Uk}(1-\bar{z}_{Uk})(1-\pi_k)(1+\frac{1}{\pi_k})\right]$$

$$= \frac{P(1-P)}{n}(\frac{n}{n-1} - \frac{1}{n-1}) - \frac{1}{(n-1)nN}\sum_{k} N_k \bar{z}_{Uk}(1-\bar{z}_{Uk})(1-\pi_k)(1+\frac{1}{\pi_k})$$

$$= \frac{P(1-P)}{n} - \frac{b}{n-1} \qquad (8.5)$$

where $b = \frac{1}{nN}\sum_{k} N_k \bar{z}_{Uk}(1-\bar{z}_{Uk})(1-\pi_k)(1+\frac{1}{\pi_k}) \doteq V_{imp,M=1} + V_{res}$

and $V_{imp,M=1}$ denotes the variance V_{imp} for $M=1$.

2.)

$$E_{DRI}\left[(1+\frac{1}{M})\frac{1}{M-1}\sum_{m=1}^{M}\left(\hat{P}_{\cdot m}-\hat{P}_{\cdot}\right)^2\right] = (1+\frac{1}{M})E_{DRI}\left[\frac{1}{M-1}\sum_{m=1}^{M}\left((\hat{P}_{\cdot m}-P)-(\hat{P}_{\cdot}-P)\right)^2\right]$$

$$= (1+\frac{1}{M})E_{DRI}\left[\frac{1}{M-1}\sum_{m=1}^{M}(\hat{P}_{\cdot m}-P)^2 - \frac{M}{M-1}(\hat{P}_{\cdot}-P)^2\right]$$

$$=(1+\frac{1}{M})\left(E_{DRI}\left[\frac{1}{M-1}\sum_{m=1}^{M}(\hat{P}_{\cdot m}-P)^2\right]-E_{DRI}\left[\frac{M}{M-1}(\hat{P}_{\cdot}-P)^2\right]\right)$$

2.1) 2.2)

Note:

2.1)

$$E_{DRI}\left[\frac{1}{M-1}\sum_{m=1}^{M}(\hat{P}_{\cdot m}-P)^2\right]$$

$$=\frac{1}{M-1}\sum_{m=1}^{M}E_{DRI}(\hat{P}_{\cdot m})^2+\mathrm{var}_{(2)DRI}(\hat{P}_{\cdot m})-2PE_{DRI}(\hat{P}_{\cdot m})+P^2$$

$$\doteq\frac{M}{M-1}\left(P^2+\frac{P(1-P)}{n}+\left(\frac{1}{nN}\sum_{k}N_k\bar{z}_{Uk}(1-\bar{z}_{Uk})(1-\pi_k)(1+\frac{1}{\pi_k})\right)-2P^2+P^2\right)$$

$$=\frac{M}{M-1}\left[\frac{P(1-P)}{n}+\frac{1}{nN}\sum_{k}N_k\bar{z}_{Uk}(1-\bar{z}_{Uk})(1-\pi_k)(1+\frac{1}{\pi_k})\right]$$

2.2)

$$E_{DRI}\left[\frac{M}{M-1}(\hat{P}_{\cdot}-P)^2\right]=\frac{M}{M-1}\left[E_{DRI}(\hat{P}_{\cdot})^2+\mathrm{var}_{(2)DRI}(\hat{P}_{\cdot})-2PE_{DRI}(\hat{P}_{\cdot})+P^2\right]$$

$$\doteq\frac{M}{M-1}\left[\frac{P(1-P)}{n}+\frac{1}{nN}\sum_{k}N_k\bar{z}_{Uk}(1-\bar{z}_{Uk})(1-\pi_k)(\frac{1}{M}+\frac{1}{\pi_k})\right]$$

It follows for 2.) that

$$E_{DRI}\left[(1+\frac{1}{M})(\frac{1}{M-1})\sum_{m=1}^{M}\left(\hat{P}_{\cdot m}-\hat{P}_{\cdot}\right)^2\right]$$

$$\doteq(1+\frac{1}{M})\,(\frac{M}{M-1})\left[\frac{1}{nN}\sum_{k}N_k\bar{z}_{Uk}(1-\bar{z}_{Uk})(1-\pi_k)\left((1+\frac{1}{\pi_k})-(\frac{1}{M}+\frac{1}{\pi_k})\right)\right]$$

$$=(1+\frac{1}{M})\frac{1}{nN}\sum_k N_k \bar{z}_{Uk}(1-\bar{z}_{Uk})(1-\pi_k) \qquad (8.6)$$

It follows that:

$$E_{DRI}(\hat{\text{var}}_{MI}(\hat{P}.)) = (1.) + (2.)$$

$$\doteq \frac{P(1-P)}{n} - \frac{b}{n-1} + (1+\frac{1}{M})\frac{1}{nN}\sum_k N_k \bar{z}_{Uk}(1-\bar{z}_{Uk})(1-\pi_k)$$

$$= \frac{P(1-P)}{n} - \frac{b}{n-1} + \frac{1}{MnN}\sum_k N_k \bar{z}_{Uk}(1-\bar{z}_{Uk})(1-\pi_k) + \frac{1}{nN}\sum_k N_k \bar{z}_{Uk}(1-\bar{z}_{Uk})(1-\pi_k) \qquad (8.7)$$

A 4.4) Proof of Result 4.6

$$\text{var}_{DRI}(\hat{P}.) \doteq \frac{1}{N^2}(E_{DR}\,\text{var}_I(q) + E_D\,\text{var}_R\,E_I(q) + \text{var}_D\,E_{RI}(q)),$$

1.) 2.) 3.)

where $q = \frac{1}{M}\sum_{m=1}^{M}\sum_{i\in s} w_i(z_{.mi} - P)$, using (4.48).

1.) $E_{DR}\,\text{var}_I(q) = E_{DR}\left(\frac{1}{M^2}\sum_{m=1}^{M}\sum_{i\in\bar{r}} w_i^2\,\text{var}(\hat{z}_{mi})\right)$, since $\hat{z}_{mi}$ independent for all m and i because

donors are selected with replacement

$$\doteq E_{DR}\left(\frac{1}{M}\sum_k\sum_{i\in B_k} I(i\in\bar{r})w_i^2\bar{z}_{rk}(1-\bar{z}_{rk})\right) = E_{DR}\left(\frac{1}{M}\sum_k \bar{z}_{rk}(1-\bar{z}_{rk})\overline{w_{rk}^2}\sum_{i\in B_k\cap\bar{r}} w_i\right)$$

$$\doteq E_D\left(\frac{1}{M}\sum_k \bar{z}_{sk}(1-\bar{z}_{sk})\overline{w_{sk}^2}\sum_{i\in B_k} w_i(1-\pi_k)\right) \doteq \frac{1}{M}\sum_k N_k\bar{z}_{Uk}(1-\bar{z}_{Uk})(1-\pi_k)\overline{w_{Uk}}$$

ignoring $\mathrm{var}_{DR}(\bar{z}_{rk})$, $E_R(\overline{w_{rk}^2}) \doteq \overline{w_{sk}^2}$ and $E_D(\overline{w_{sk}^2}) \doteq \overline{w_{Uk}}$, since

$$E_D\left(\frac{1}{\sum_{i\in B_{Uk}} w_i I(i\in s)} \sum_{i\in B_{Uk}} w_i^2 I(i\in s)\right) \doteq \frac{1}{N_k}\sum_{i\in B_{Uk}} w_i = \overline{w_{Uk}}\,.$$

2.) $E_D\,\mathrm{var}_R\,E_I\left(\frac{1}{M}\sum_{m=1}^{M}\sum_{i\in s} w_i(z_{\cdot mi} - P)\right)$

$$= E_D\,\mathrm{var}_R\left(\sum_k \sum_{i\in B_k} I(i\in r) w_i (z_i - P) + (1 - I(i\in r)) w_i(\bar{z}_{rk} - P)\right)$$

$$= E_D\,\mathrm{var}_R\left(\sum_k \sum_{i\in B_k} I(i\in r) w_i (z_i - \bar{z}_{rk}) + w_i(\bar{z}_{rk} - P)\right)$$

Note: $\sum_{i\in B_k} I(i\in r) w_i \bar{z}_{rk} = \sum_{i\in B_k} I(i\in r) w_i z_i$

$$= E_D\,\mathrm{var}_R\left(\sum_k \sum_{i\in B_k} w_i(\bar{z}_{rk} - P)\right) \qquad (8.8)$$

$$= E_D\,\mathrm{var}_R\left(\sum_k \sum_{i\in B_k} w_i \bar{z}_{rk}\right) = E_D\left(\sum_k (\mathrm{var}_R(\bar{z}_{rk})\,(\sum_{i\in B_k} w_i)^2)\right)$$

Note (see proof below in (8.9)): $\mathrm{var}_R(\bar{z}_{rk}) \doteq (\sum_{i\in B_k} w_i)^{-2} \sum_{i\in B_k} w_i^2 (z_i - \bar{z}_{sk})^2 \left(\frac{1}{\pi_k} - 1\right)$

$$\doteq E_D\left(\sum_k \sum_{i\in B_k} \left(\frac{1}{\pi_k} - 1\right) w_i^2 (z_i - \bar{z}_{sk})^2\right)$$

for calculating the expectation we basically need to obtain $E_D(\bar{z}_{sk} \sum_{i\in B_k} w_i^2 z_i)$ and

$E_D(\bar{z}_{sk}^2 \sum_{i\in B_k} w_i^2)$. Using Taylor series expansion it follows:

$$\doteq \sum_k \sum_{i\in B_{Uk}} \left(\frac{1}{\pi_k} - 1\right) w_i z_i^2 - 2\sum_k \sum_{i\in B_{Uk}} \left(\frac{1}{\pi_k} - 1\right) w_i z_i \bar{z}_{Uk} + \sum_k \sum_{i\in B_{Uk}} \left(\frac{1}{\pi_k} - 1\right) \bar{z}_{Uk}^2 w_i$$

$$= \sum_k \sum_{i \in B_{Uk}} \left(\frac{1}{\pi_k} - 1\right) w_i \, (z_i - \bar{z}_{Uk})^2$$

We still need to show:

Using Taylor series expansion $\bar{z}_{rk}$ can be approximated similarly to (4.47) since the number of respondents in each class B_k is assumed large. It follows for the variance

$$\mathrm{var}_R(\bar{z}_{rk}) = \mathrm{var}_R\Big(\sum_{i \in B_k \cap r} w_i \, z_i \Big/ \sum_{i \in B_k \cap r} w_i\Big) \doteq \frac{1}{\left(\pi_k \sum_{i \in B_k} w_i\right)^2} \mathrm{var}_R\left(\sum_{i \in B_k \cap r} w_i \, z_i - \sum_{i \in B_k \cap r} w_i \, \bar{z}_{sk}\right)$$

$$= \frac{1}{\left(\pi_k \sum_{i \in B_k} w_i\right)^2} \mathrm{var}_R\left(\sum_{i \in B_k} w_i \, (z_i - \bar{z}_{sk}) I(i \in r)\right)$$

$$= \frac{1}{\left(\pi_k \sum_{i \in B_k} w_i\right)^2} \sum_{i \in B_k} w_i^2 (z_i - \bar{z}_{sk})^2 \, \pi_k (1 - \pi_k)$$

since $I(i \in r)$ independent of $I(j \in r) \;\forall\; i \neq j$ under equal probability sampling

$$= \frac{1}{\left(\sum_{i \in B_k} w_i\right)^2} \sum_{i \in B_k} w_i^2 (z_i - \bar{z}_{sk})^2 \left(\frac{1}{\pi_k} - 1\right) \tag{8.9}$$

3.) $\mathrm{var}_D \, E_R E_I \left(\frac{1}{M} \sum_{m=1}^{M} \sum_{i \in s} w_i (z_{.mi} - P)\right) = \mathrm{var}_D \, E_R \left(\sum_k \sum_{i \in B_k} w_i (\bar{z}_{rk} - P)\right)$

compare to 2.) in this proof, equation (8.8)

$$= \mathrm{var}_D \left(\sum_k \sum_{i \in B_k} w_i (\bar{z}_{sk} - P)\right) = \mathrm{var}_D \left(\sum_{i \in s} w_i (z_i - P)\right) = \mathrm{var}_D \left(\sum_{i \in s} w_i a_i\right)$$

where $a_i = z_i - P$, and $\bar{z}_{sk} \sum_{i \in B_k} w_i = \sum_{i \in B_k} w_i z_i$

$= \mathrm{var}_D \left(\sum_{i \in U} w_i a_i I(i \in s)\right) \doteq \sum_{i \in U} (w_i a_i)^2 \, \mathrm{var}_D (I(i \in s))$, since the covariance terms are assumed

negligible. (If one were not willing to make this assumption the approach used by Särndal, Swensson and Wretman (1992, section 2.8, result 2.8.1) can be applied based on joint inclusion probabilities).

$$= \sum_{i \in U} (w_i a_i)^2 \frac{1}{w_i}(1 - \frac{1}{w_i}) = \sum_{i \in U} w_i a_i^{\,2}(1 - \frac{1}{w_i}) = \sum_{i \in U} a_i^{\,2}(w_i - 1) = \sum_{i \in U} w_i a_i^{\,2} - \sum_{i \in U} a_i^{\,2}$$

$$= \sum_{i \in U} w_i a_i^{\,2} - \sum_{i \in U} a_i = \sum_{i \in U} w_i (z_i - P)^2 - \sum_{i \in U} (z_i - P)$$

$$= \sum_{i \in U} w_i (z_i - P)^2 \text{ , since } \sum_{i \in U} z_i - NP = 0$$

It follows:

$\text{var}_{DRI}(\hat{P}.) \doteq \frac{1}{N^2}\left((1.) + (2.) + (3.) \right)$, which gives the formula in result 4.6.

❑

A 4.5) Proof of Result 4.8

We can write

$$\mathrm{var}_{DRI}(\hat{P}.) \doteq \frac{1}{N^2}\left\{\sum_{i\in U} w_i (z_i - P)^2 + \frac{1}{M}\sum_k N_k \bar{z}_{Uk}(1-\bar{z}_{Uk})(1-\pi_k)\,\overline{w_{Uk}}\right.$$

$$\left. + \sum_k \sum_{i\in B_k} (\frac{1}{\pi_k} - 1) w_i (z_i - \bar{z}_{Uk})^2 \right\}$$

$$= \lambda \{A+B+C\}.$$

Since we have to show $E_{DRI}(\hat{\mathrm{var}}_{DRI}(\hat{P}.)) \doteq \mathrm{var}_{DRI}(\hat{P}.)$, we need to find approximately unbiased estimators of the four components λ, A, B and C, using Taylor approximations. We have

a.) An approximately unbiased estimator of λ is $\hat{\lambda} = \frac{1}{(\sum_{i\in s} w_i)^2}$.

$E_D(\hat{\lambda}) = \left(E_D((\sum_{i\in s} w_i)^2)\right)^{-1} \doteq \left((E_D(\sum_{i\in s} w_i))^2\right)^{-1} = \frac{1}{N^2}$, since $E_D(\sum_{i\in s} w_i) = N$ (see proof of result 3.2) and using Taylor approximation as in equation (3.14) based on the function $g(\bar{x}_s) = \bar{x}_s^{\,2}$.

b.) $\hat{A} = \frac{1}{M}\sum_{m=1}^{M}\sum_k \sum_{i\in B_k} \frac{1}{\hat{\pi}_k} w_i^2 (z_{.mi} - \hat{P}.)^2 - \sum_k \sum_{i\in B_k} (\frac{1}{\hat{\pi}_k} - 1) w_i^2 \left((\bar{z}_{.sk} - \hat{P}.)^2 + \bar{z}_{.sk}(1-\bar{z}_{.sk})\right)$

We need to show that $E_{DRI}(\hat{A}) = A$. However, for the purpose of this proof we substitute $\hat{P}.$ by P, since

$$\sum_{i\in s} w_i^2 (z_{.mi} - \hat{P}.)^2 = \sum_{i\in s} w_i^2 ((z_{.mi} - P) - (\hat{P}. - P))^2$$

$$= \sum_{i\in s} w_i^2 (z_{.mi} - P)^2 - 2\sum_{i\in s} w_i^2 (z_{.mi} - P)(\hat{P}. - P) + \sum_{i\in s} w_i^2 (\hat{P}. - P)^2$$

$$\doteq \sum_{i\in s} w_i^2 (z_{.mi} - P)^2 \quad \text{since } (\hat{P}. - P) \text{ is of smaller order.}$$

$E_{DRI}(\hat{A})$

$$\doteq E_{DR}\left(\sum_k \sum_{i\in B_k} I(i\in r)\frac{1}{\hat{\pi}_k} w_i^2 (z_i - P)^2 + I(i\in \bar{r})\frac{1}{\hat{\pi}_k} w_i^2 \left((\bar{z}_{rk} - P)^2 + \bar{z}_{rk}(1-\bar{z}_{rk})\right)\right)$$

$$-E_D\left(\sum_k \sum_{i\in B_k} (\frac{1}{\pi_k} - 1) w_i^2 \left((\bar{z}_{sk} - P)^2 + \bar{z}_{sk}(1-\bar{z}_{sk})\right)\right),$$

since $\mathrm{var}_I(\hat{z}_{mi} - P) = \mathrm{var}_I(\hat{z}_{mi}) = \bar{z}_{rk}(1-\bar{z}_{rk})$.

Using Taylor series approximation it follows

$$\doteq E_D\left(\sum_k \sum_{i\in B_k} w_i^2 (z_i - P)^2 + (\frac{1}{\pi_k} - 1) w_i^2 \left((\bar{z}_{sk} - P)^2 + \bar{z}_{sk}(1-\bar{z}_{sk})\right)\right)$$

$$-E_D\left(\sum_k \sum_{i\in B_k} (\frac{1}{\pi_k} - 1) w_i^2 \left((\bar{z}_{sk} - P)^2 + \bar{z}_{sk}(1-\bar{z}_{sk})\right)\right)$$

$$\doteq \sum_k \sum_{i\in B_{Uk}} w_i (z_i - P)^2 + (\frac{1}{\pi_k} - 1) w_i \left((\bar{z}_{Uk} - P)^2 + \bar{z}_{Uk}(1-\bar{z}_{Uk})\right)$$

$$-\sum_k \sum_{i\in B_{Uk}} (\frac{1}{\pi_k} - 1) w_i \left((\bar{z}_{Uk} - P)^2 + \bar{z}_{Uk}(1-\bar{z}_{Uk})\right)$$

$$= \sum_{i\in s} w_i (z_i - P)^2 = \mathrm{A}$$

c.) An approximately unbiased estimator of B is $\hat{B} = \frac{1}{M}\sum_k (\bar{z}_{\cdot sk}(1-\bar{z}_{\cdot sk})(1-\hat{\pi}_k)\sum_{i\in B_k} w_i^2)$.

$$E_{DRI}(\hat{B}) \doteq \frac{1}{M}\sum_k (\bar{z}_{Uk}(1-\bar{z}_{Uk})(1-\pi_k)\sum_{i\in B_{Uk}} w_i) = \frac{1}{M}\sum_k N_k \bar{z}_{Uk}(1-\bar{z}_{Uk})(1-\pi_k)\overline{w_{Uk}} = B.$$

d.) $$\hat{C} = \frac{1}{M}\sum_{m=1}^{M}\sum_k \sum_{i\in B_k} \frac{1}{\hat{\pi}_k}(\frac{1}{\hat{\pi}_k} - 1) w_i^2 (z_{\cdot mi} - \bar{z}_{\cdot sk})^2 - \sum_k \sum_{i\in B_k} (\frac{1}{\hat{\pi}_k} - 1)^2 w_i^2 \bar{z}_{\cdot sk}(1-\bar{z}_{\cdot sk})$$

For the purpose of this proof we substitute $\bar{z}_{.sk}$ by $\bar{z}_{Uk}$ similar to part b.) where we substituted $\hat{P}.$ by P.

$$E_{DRI}(\hat{C}) \doteq E_{DR}\left(\sum_k \frac{1}{\hat{\pi}_k}\left(\frac{1}{\hat{\pi}_k}-1\right)\sum_{i\in B_k} I(i\in r)w_i^2(z_i-\bar{z}_{Uk})^2 + I(i\in \bar{r})w_i^2\left((\bar{z}_{rk}-\bar{z}_{Uk})^2+\bar{z}_{rk}(1-\bar{z}_{rk})\right)\right)$$

$$-E_D\left(\sum_k \sum_{i\in B_k}\left(\frac{1}{\pi_k}-1\right)^2 w_i^2 \bar{z}_{sk}(1-\bar{z}_{sk})\right)$$

using Taylor series approximation it follows:

$$\doteq \sum_k \sum_{i\in B_{Uk}}\left(\frac{1}{\pi_k}-1\right)w_i(z_i-\bar{z}_{Uk})^2 + \left(\frac{1}{\pi_k}-1\right)^2 w_i\left((\bar{z}_{Uk}-\bar{z}_{Uk})^2+\bar{z}_{Uk}(1-\bar{z}_{Uk})\right)$$

$$-\sum_k \sum_{i\in B_{Uk}}\left(\frac{1}{\pi_k}-1\right)^2 w_i \bar{z}_{Uk}(1-\bar{z}_{Uk})$$

$$=\sum_k \sum_{i\in B_{Uk}}\left(\frac{1}{\pi_k}-1\right)w_i(z_i-\bar{z}_{Uk})^2$$

It follows: $\hat{\lambda}\{\hat{A}+\hat{B}+\hat{C}\} = \hat{\text{var}}_{DRI}(\hat{P}.)$

❑

Appendix: Chapter 6

A 6.1) Estimated Linear Regression for the Additive Model, based on March-May 2000 quarter:

$$x_i = \hat{\alpha} + \hat{\beta} y_i + \hat{\varepsilon}_i,$$

where X is the derived and Y the direct variable. Type 1 outliers are excluded.

```
Table of Coefficients:
                    Value Std. Error  t value Pr(>|t|)
      (Intercept)  0.8970    0.0508  17.6520   0.0000
   direct variable 0.8873    0.0073 120.7936   0.0000


Analysis of Variance Table:
                    Df Sum of Sq  Mean Sq F Value  Pr(F)
   direct variable   1  52520.03 52520.03 14591.1      0
         Residuals 6824  24562.69     3.60

Multiple R-Squared: 0.6813
F-statistic: 14590 on 1 and 6824 degrees of freedom, the p-value is 0
```

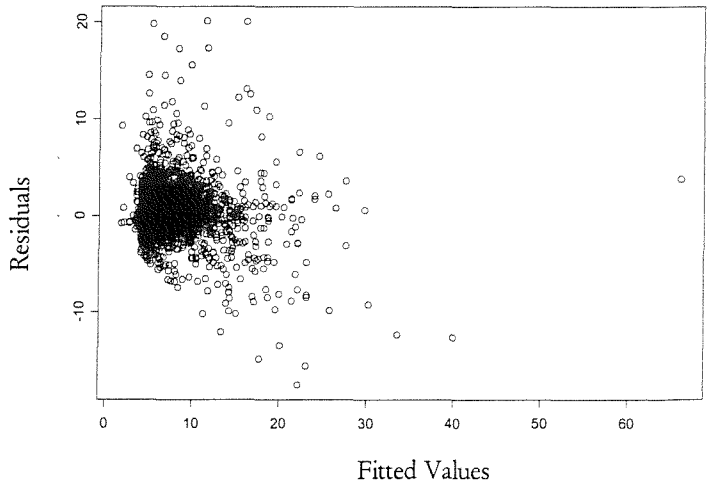

Figure A 6.1.1: Plot of residuals vs fitted values.

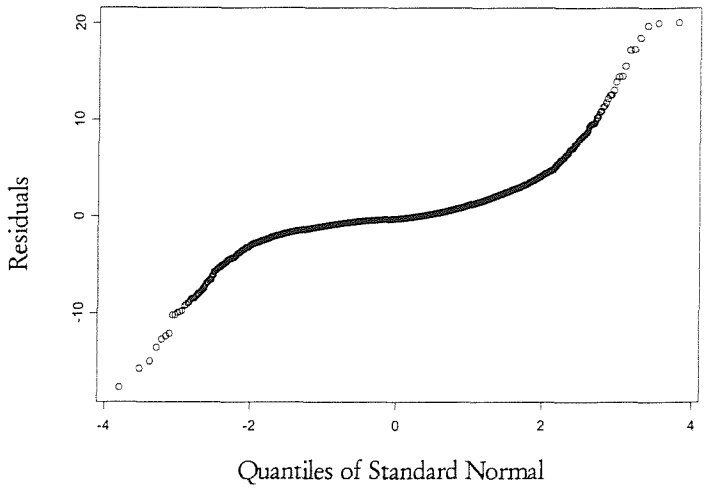

Figure A 6.1.2: Normal probability plot

A 6.2) Estimated Linear Regression for the Multiplicative Model, based on March-May 2000 quarter:

$$\ln(x_i) = \hat{\alpha} + \hat{\beta}\ln(y_i) + \hat{\varepsilon}_i .$$

Type 1 outliers are excluded.

```
Table of Coefficients:
                Value Std. Error  t value Pr(>|t|)
    (Intercept) 0.0921   0.0162     5.6728   0.0000
         ln(Y)  0.9511   0.0091   104.0982   0.0000

Analysis of Variance Table:
               Df Sum of Sq  Mean Sq  F Value Pr(F)
         ln(Y)  1  884.9734 884.9734 10836.44     0
    Residuals 6824  557.2919   0.0817

Multiple R-Squared: 0.6136
F-statistic: 10840 on 1 and 6824 degrees of freedom, the p-value is 0
```

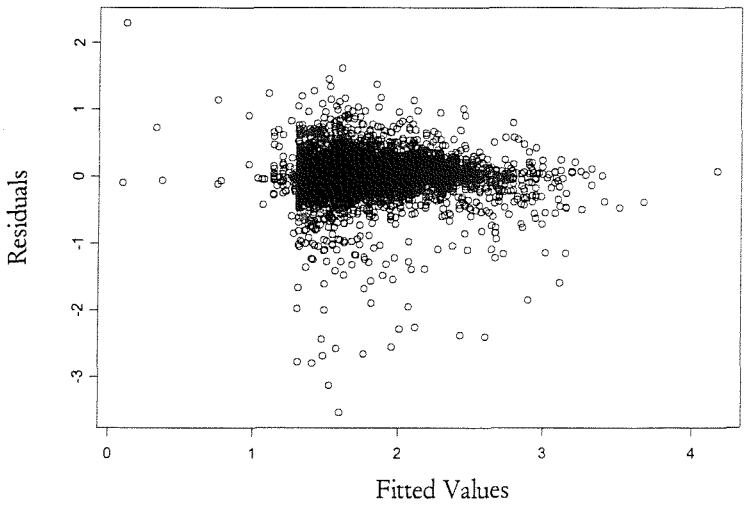

Figure A 6.2.1: Plot of residuals vs fitted values

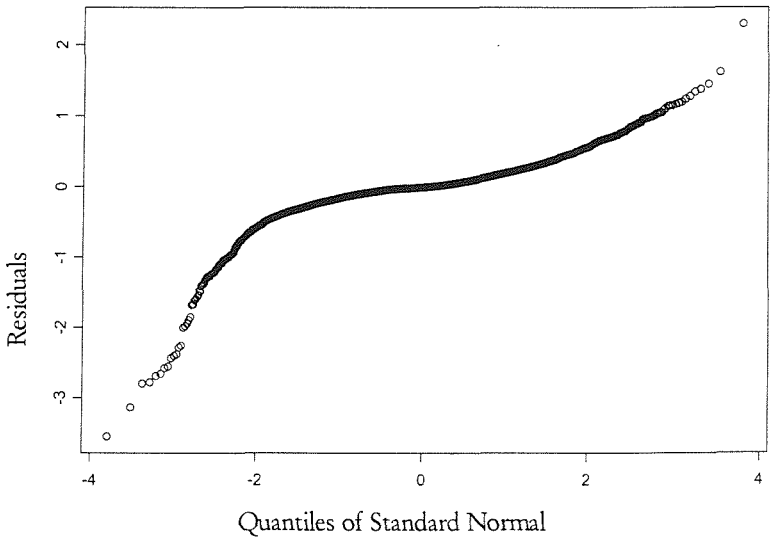

Figure A 6.2.2: Normal probability plot

A 6.3) Estimated Linear Regression to Analyze the Dependence of Measurement Error on other Covariates, based on March-May 2000 quarter:

$$d_i = \ln(x_i) - \ln(y_i) = g(\ln(y_i)) + \mu_i\hat{\beta} + \hat{\varepsilon}_i,$$

including some two-way interactions and excluding type 2 outliers.

Table of Coefficients:

	Value	Std. Error	t value	Pr(>|t|)
(Intercept)	0.0251	0.0579	0.4340	0.6643
ln(Y)	0.1245	0.0487	2.5592	0.0105
ln(Y)squared	-0.0556	0.0123	-4.5130	0.0000
FEMALE	-0.0196	0.0088	-2.2240	0.0262
ADDTBP	0.0802	0.0061	13.1645	0.0000
USESLP	-0.0072	0.0019	-3.7865	0.0002
LTWK	1.3541	0.1728	7.8365	0.0000
EVEROTcat	-0.1138	0.0280	-4.0644	0.0000
SOCcat1	-0.0541	0.0187	-2.8967	0.0038
SOCcat2	-0.0226	0.0145	-1.5598	0.1188
SOCcat3	-0.0495	0.0131	-3.7811	0.0002
SOCcat4	-0.0518	0.0142	-3.6534	0.0003
SOCcat5	-0.0582	0.0135	-4.2973	0.0000
SOCcat6	-0.0386	0.0142	-2.7071	0.0068
SOCcat7	-0.0320	0.0140	-2.2846	0.0224
SOCcat8	-0.0490	0.0139	-3.5223	0.0004
USGRS	-0.0509	0.0063	-8.0775	0.0000
MARRIED	0.0154	0.0052	2.9785	0.0029
INDcat1	0.1053	0.0397	2.6560	0.0079
INDcat2	0.0717	0.0305	2.3542	0.0186
INDcat3	0.1039	0.0320	3.2490	0.0012
INDcat4	0.0316	0.0304	1.0385	0.2991
INDcat5	0.0752	0.0314	2.3984	0.0165
INDcat6	0.0489	0.0311	1.5732	0.1157
INDcat7	0.0631	0.0306	2.0659	0.0389
INDcat8	0.0546	0.0318	1.7181	0.0858
ln(Y):LTWK	-0.4910	0.1100	-4.4632	0.0000
ln(Y):EVEROTcat	0.0432	0.0145	2.9726	0.0030
FEMALE:EVEROTcat	-0.0283	0.0109	-2.5995	0.0094
ADDTBP:LTWK	-0.3586	0.1510	-2.3753	0.0176

Multiple R-Squared: 0.1157
F-statistic: 30.18 on 29 and 6687 degrees of freedom, the p-value is 0

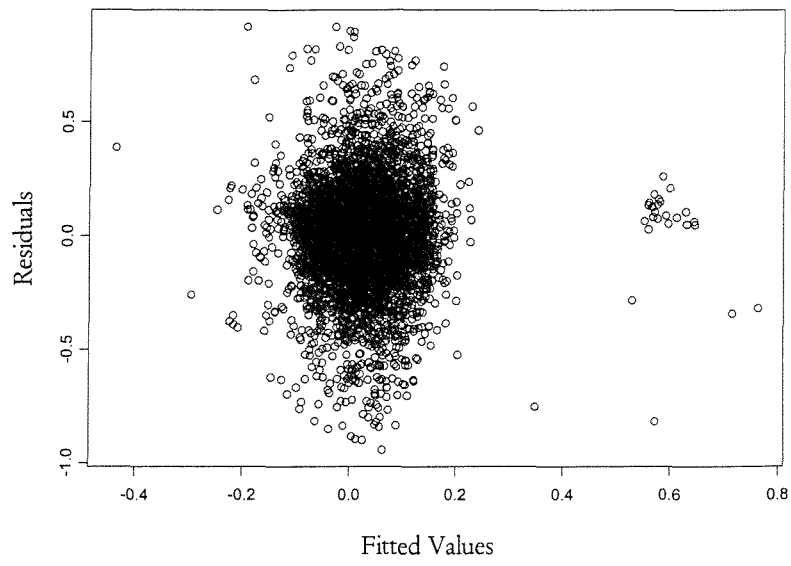

Figure A 6.3.1: Plot of the residuals vs fitted values

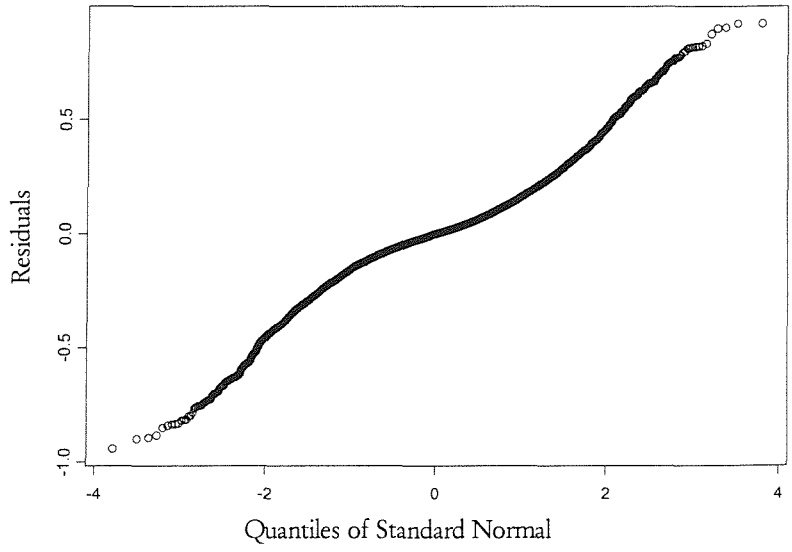

Figure A 6.3.2: Normal probability plot

A 6.4) Estimated Linear Regression Model to Predict the Direct Variable, based on respondents only, based on March-May 2000 quarter:

$$\ln(y_i) = \mu_i \hat{\beta} + \hat{\varepsilon}_i,$$

where μ_i is a row-vector of functions of the covariates W.

Table of Coefficients:

	Value	Std. Error	t value	Pr(>|t|)
(Intercept)	1.6622	0.0518	32.0635	0.0000
sampleh$SOCcat1	0.0971	0.0270	3.5965	0.0003
sampleh$SOCcat2	-0.0162	0.0207	-0.7853	0.4323
sampleh$SOCcat3	-0.2437	0.0182	-13.3585	0.0000
sampleh$SOCcat4	-0.2044	0.0183	-11.1570	0.0000
sampleh$SOCcat5	-0.3513	0.0189	-18.5804	0.0000
sampleh$SOCcat6	-0.3180	0.0231	-13.7905	0.0000
sampleh$SOCcat7	-0.3487	0.0181	-19.2933	0.0000
sampleh$SOCcat8	-0.4374	0.0203	-21.5916	0.0000
sampleh$PT	-0.2387	0.0339	-7.0368	0.0000
sampleh$LTWK	-0.0072	0.0441	-0.1640	0.8697
sampleh$Qcat1	-0.0654	0.0169	-3.8711	0.0001
sampleh$Qcat2	-0.1752	0.0152	-11.5051	0.0000
sampleh$Qcat3	-0.2199	0.0151	-14.5137	0.0000
sampleh$Qcat4	-0.2694	0.0160	-16.8865	0.0000
sampleh$Qcat5	-0.2999	0.0164	-18.2484	0.0000
sampleh$AGE	0.0167	0.0017	9.8969	0.0000
sampleh$AGEsquared	-0.0002	0.0000	-10.2095	0.0000
sampleh$EMPMONcat1	0.0077	0.0102	0.7572	0.4489
sampleh$EMPMONcat2	0.0470	0.0091	5.1927	0.0000
sampleh$EMPMONcat3	0.1417	0.0084	16.8429	0.0000
sampleh$HOH	0.0563	0.0087	6.4545	0.0000
sampleh$MARRIED	0.0452	0.0074	6.1297	0.0000
sampleh$SIZE	0.0752	0.0066	11.3749	0.0000
sampleh$INDcat1	0.2136	0.0468	4.5659	0.0000
sampleh$INDcat2	0.1425	0.0361	3.9540	0.0001
sampleh$INDcat3	0.1788	0.0378	4.7273	0.0000
sampleh$INDcat4	0.0236	0.0361	0.6549	0.5126
sampleh$INDcat5	0.1186	0.0372	3.1926	0.0014
sampleh$INDcat6	0.1745	0.0367	4.7476	0.0000
sampleh$INDcat7	0.1243	0.0363	3.4287	0.0006
sampleh$INDcat8	0.0247	0.0377	0.6549	0.5125
sampleh$REGcat1	-0.0142	0.0165	-0.8628	0.3883
sampleh$REGcat2	0.0006	0.0165	0.0381	0.9696
sampleh$REGcat3	-0.0113	0.0176	-0.6407	0.5217
sampleh$REGcat4	0.0065	0.0168	0.3882	0.6979
sampleh$REGcat5	0.0690	0.0171	4.0405	0.0001
sampleh$REGcat6	0.1659	0.0177	9.3473	0.0000
sampleh$REGcat7	0.0983	0.0159	6.1842	0.0000
sampleh$REGcat8	0.0374	0.0167	2.2379	0.0253
sampleh$REGcat9	0.0022	0.0192	0.1164	0.9073
sampleh$REGcat10	0.0157	0.0164	0.9540	0.3401
sampleh$REGcat11	-0.0315	0.0236	-1.3352	0.1818
sampleh$FEMALE	-0.0893	0.0096	-9.3332	0.0000
sampleh$USGRS	-0.0096	0.0071	-1.3564	0.1750
sampleh$SOCcat1:sampleh$PT	0.6417	0.0476	13.4930	0.0000
sampleh$SOCcat2:sampleh$PT	0.1774	0.0400	4.4324	0.0000
sampleh$SOCcat3:sampleh$PT	0.2031	0.0369	5.4959	0.0000
sampleh$SOCcat4:sampleh$PT	0.0050	0.0509	0.0975	0.9223
sampleh$SOCcat5:sampleh$PT	0.1781	0.0367	4.8562	0.0000
sampleh$SOCcat6:sampleh$PT	0.1638	0.0390	4.1968	0.0000
sampleh$SOCcat7:sampleh$PT	0.1377	0.0421	3.2677	0.0011
sampleh$SOCcat8:sampleh$PT	0.2171	0.0379	5.7276	0.0000

```
Residual standard error: 0.2459 on 6750 degrees of freedom
Multiple R-Squared: 0.5847
F-statistic: 182.8 on 52 and 6750 degrees of freedom, the p-value is 0
```

Analysis of Variance Table

	Df	Sum of Sq	Mean Sq	F Value	Pr(F)
sampleh$SOCcat	8	389.0176	48.62719	804.2684	0.0000000
sampleh$PT	1	22.7908	22.79076	376.9472	0.0000000
sampleh$LTWK	1	0.0452	0.04516	0.7470	0.3874671
sampleh$Qcat	5	32.7353	6.54707	108.2851	0.0000000
sampleh$AGE	1	13.9325	13.93246	230.4356	0.0000000
sampleh$AGEsquared	1	14.2176	14.21759	235.1516	0.0000000
sampleh$EMPMONcat	3	21.6715	7.22382	119.4782	0.0000000
sampleh$HOH	1	9.3899	9.38994	155.3048	0.0000000
sampleh$MARRIED	1	5.3534	5.35338	88.5421	0.0000000
sampleh$SIZE	1	10.7316	10.73163	177.4955	0.0000000
sampleh$INDcat	8	16.4771	2.05964	34.0654	0.0000000
sampleh$REGcat	11	17.6820	1.60745	26.5865	0.0000000
sampleh$FEMALE	1	5.4603	5.46031	90.3106	0.0000000
sampleh$USGRS	1	0.2304	0.23040	3.8108	0.0509659
sampleh$SOCcat:sampleh$PT	8	14.8667	1.85833	30.7359	0.0000000
Residuals	6750	408.1145	0.06046		

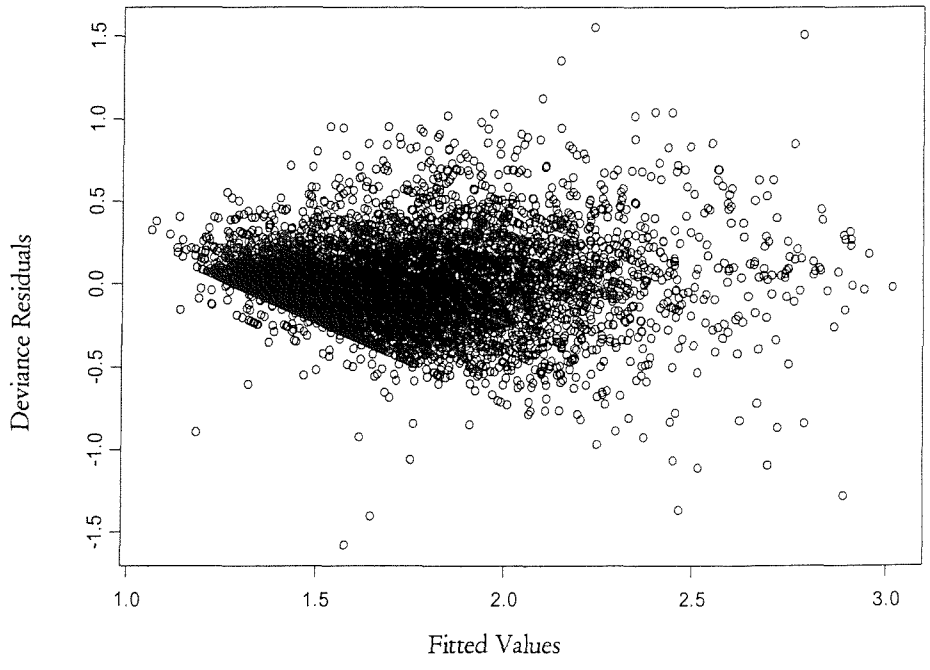

Figure A 6.4.1: Plot of residuals versus fitted values.

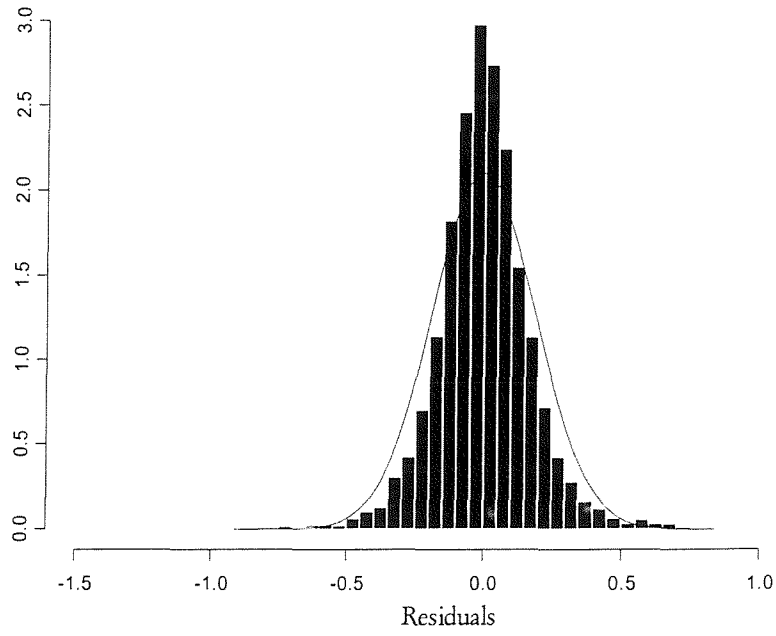

Figure A 6.4.2: Histogram of the residuals with normal density curve.

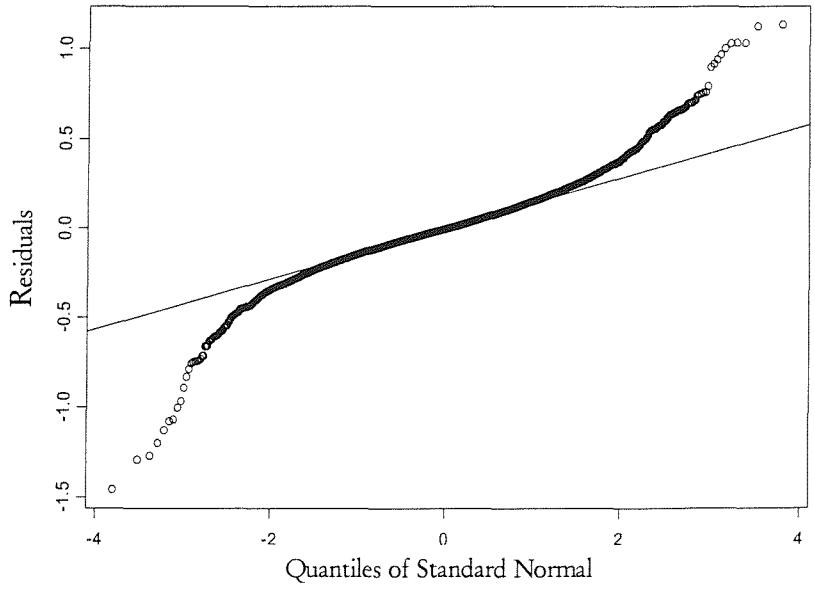

Figure A 6.4.3: Normal probability plot.

A 6.5) Selected Time Series Plots for Data Augmentation based on the CME Assumption using Random Regression Imputation

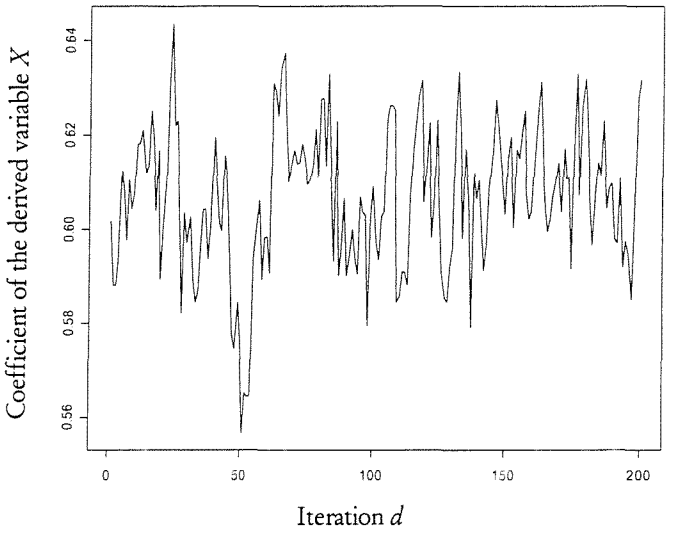

Figure A 6.5.1: Time series plot for the coefficient of the derived variable X in the imputation model over the first 200 iterations under DA-CME reg imp.

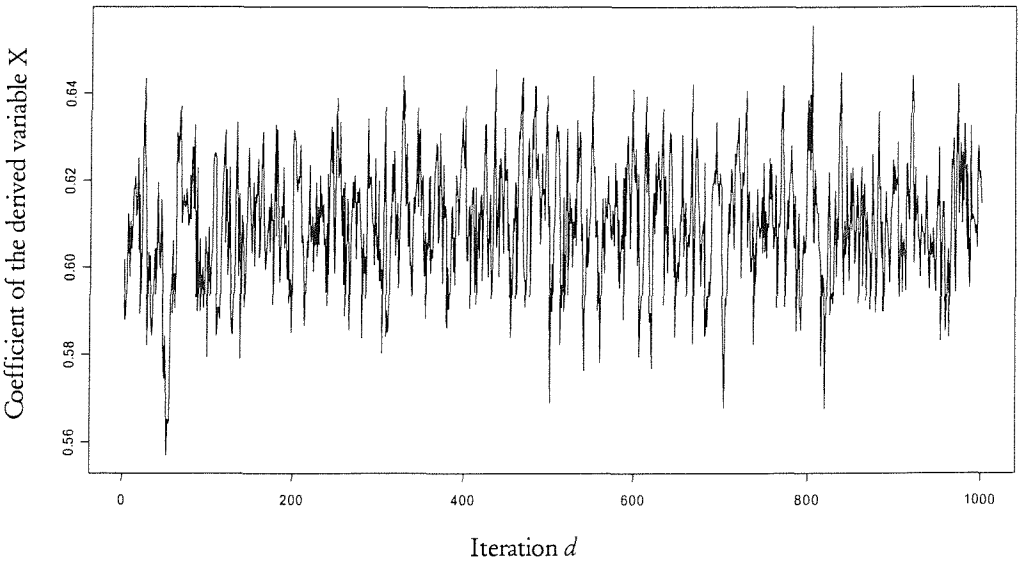

Figure A 6.5.2: Time series plot for the coefficient of the derived variable X in the imputation model over the first 1000 iterations under DA-CME reg imp.

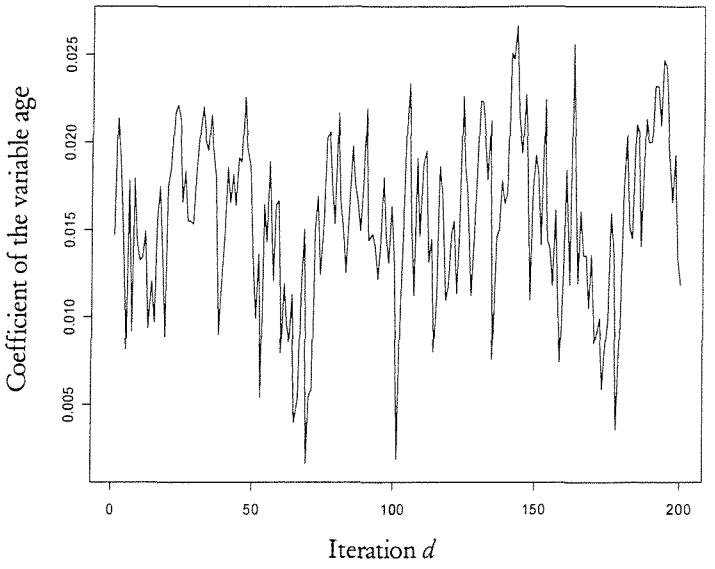

Figure A 6.5.3: Time series plot for the coefficient of the variable age in the imputation model over the first 200 iterations under DA-CME reg imp.

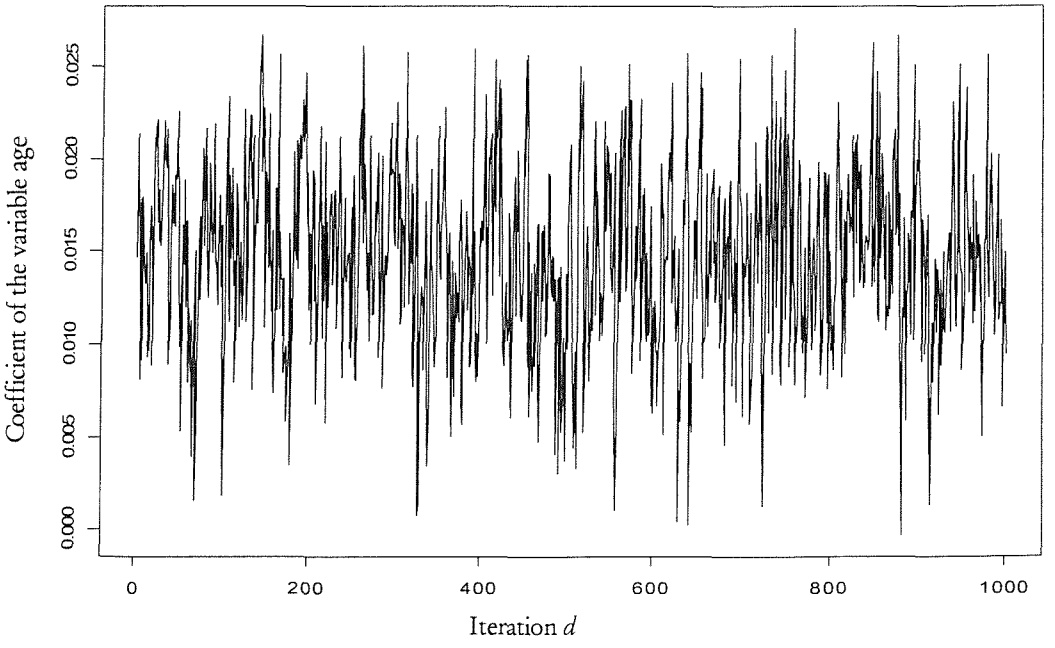

Figure A 6.5.4: Time series plot for the coefficient of the variable age in the imputation model over the first 1000 iterations under DA-CME reg imp.

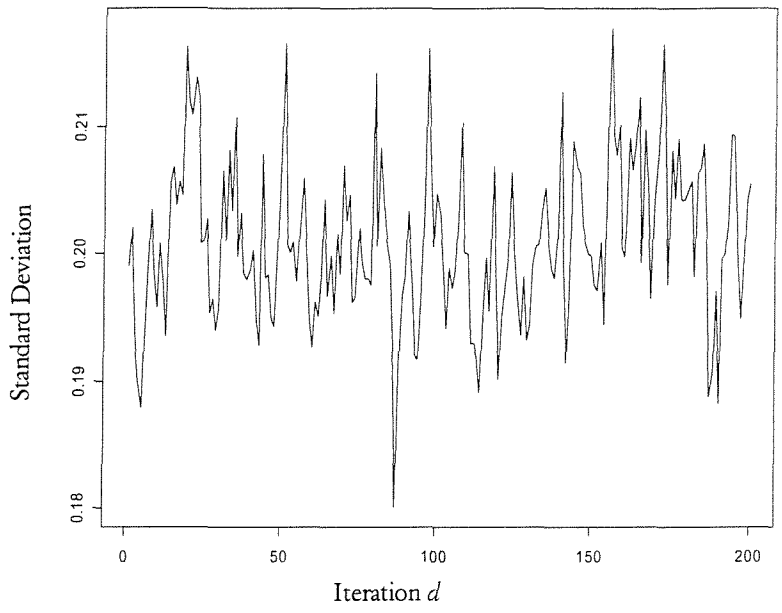

Figure A 6.5.5: Time series plot for the standard deviation of the residuals for the imputation model over the first 200 iterations under DA-CME reg imp.

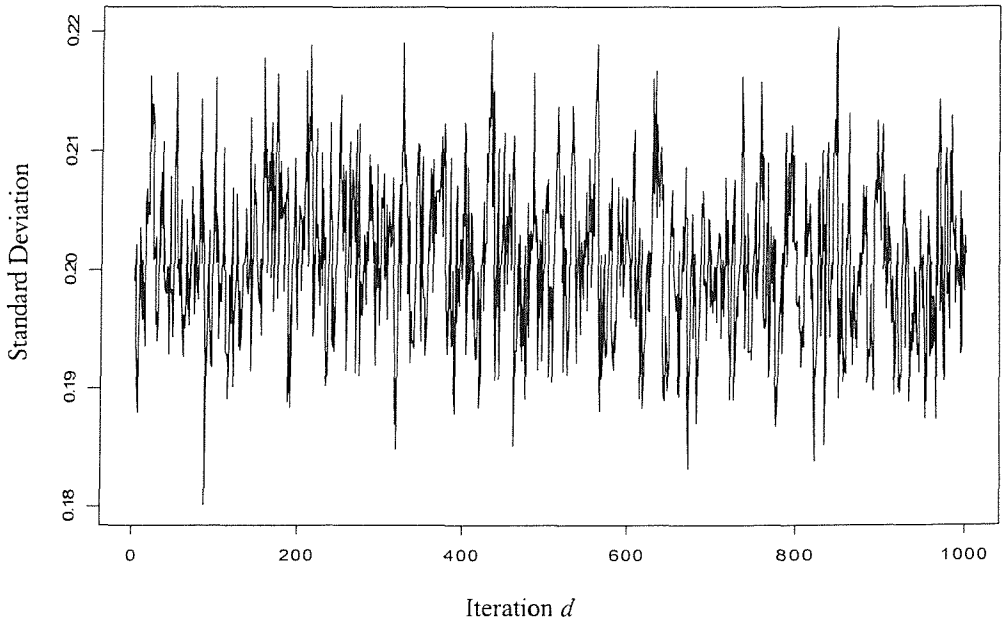

Figure A 6.5.6: Time series plot for the standard deviation of the residuals for the imputation model over the first 1000 iterations under DA-CME reg imp.

A 6.6) Selected Time Series Plots for Data Augmentation based on the CME Assumption using Nearest Neighbour Imputation, Q=10.

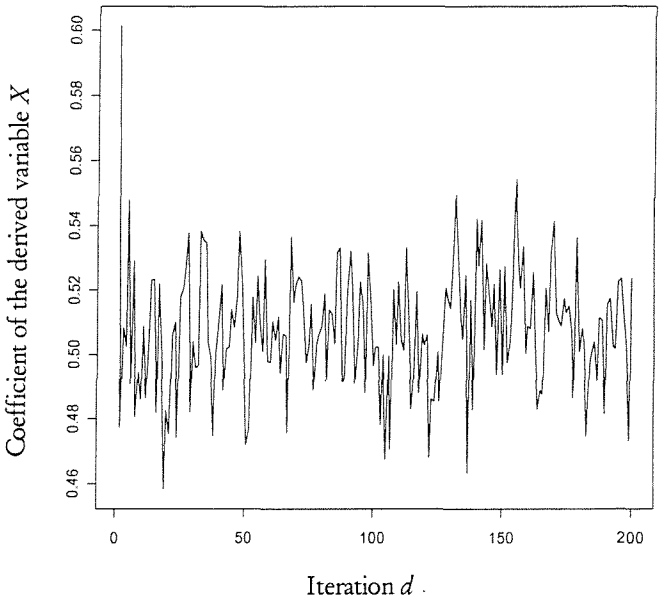

Figure A 6.6.1: Time series plot for the coefficient of the derived variable in the imputation model over the first 200 iterations under DA-CME NN Q=10.

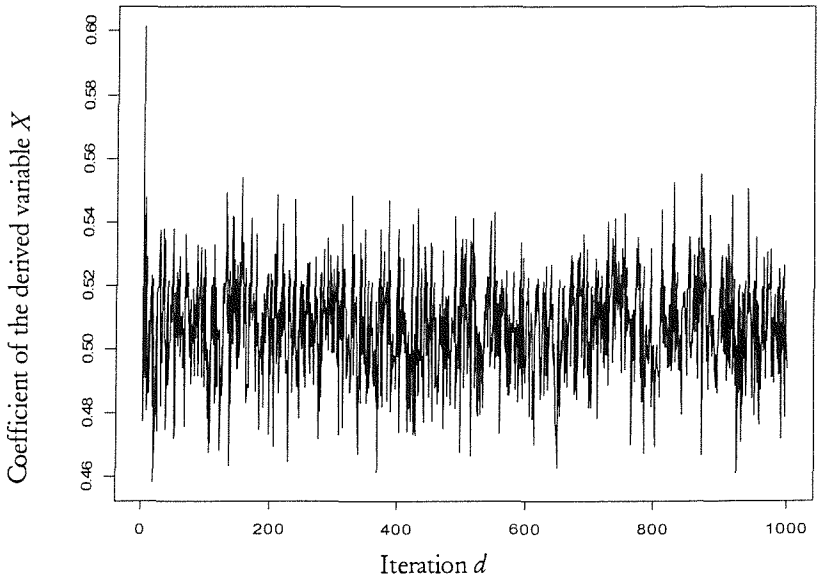

Figure A 6.6.2: Time series plot for the coefficient of the derived variable in the imputation model over the first 1000 iterations under DA-CME NN Q=10.

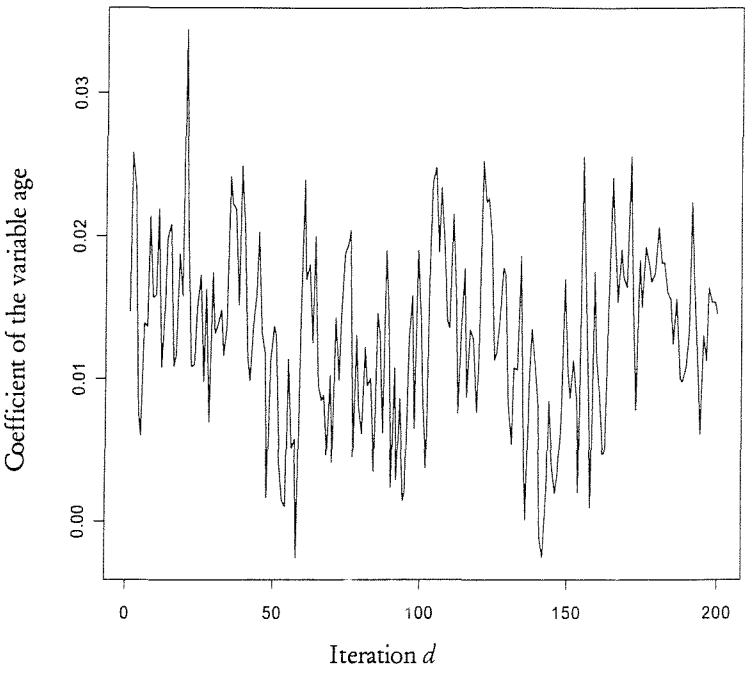

Figure A 6.6.3: Time series plot for the coefficient of the variable age in the imputation model over the first 200 iterations under DA-CME NN Q=10.

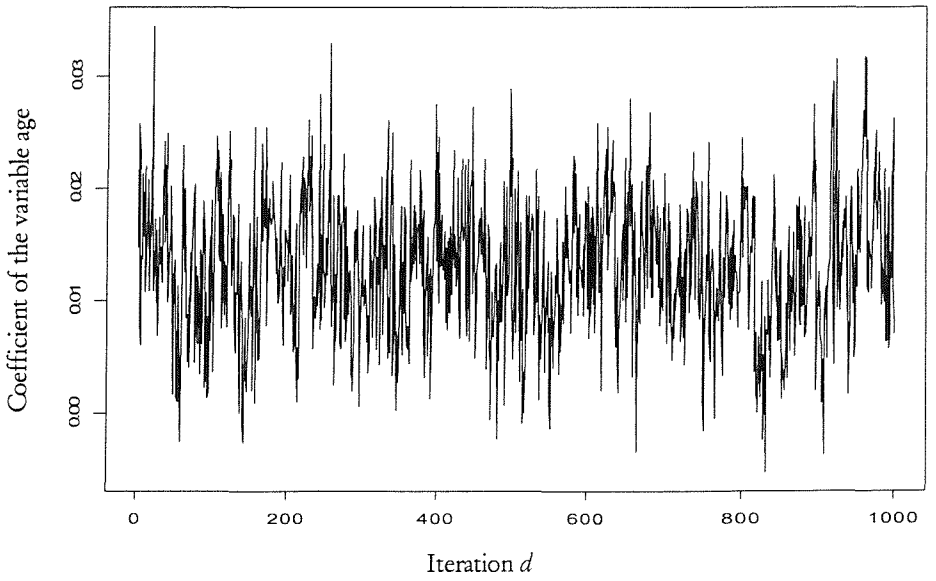

Figure A 6.6.4: Time series plot for the coefficient of the variable age in the imputation model over the first 1000 iterations under DA-CME NN Q=10.

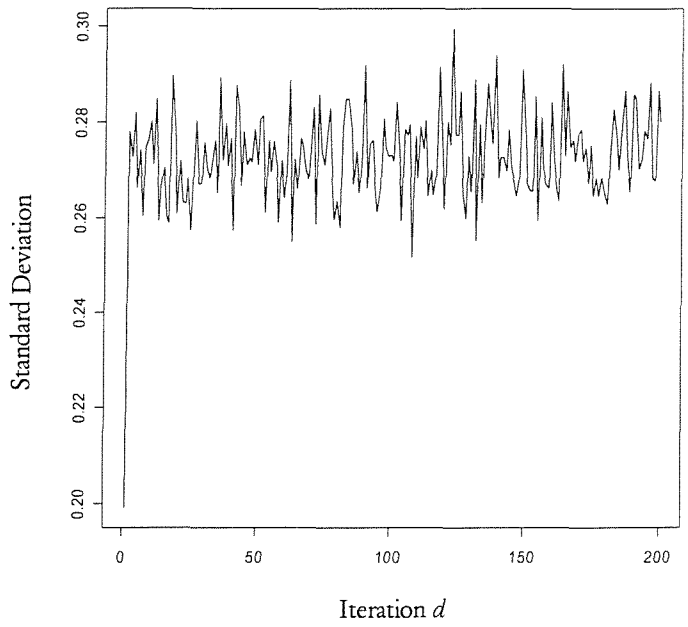

Figure A 6.6.5: Time series plot for the standard deviation of the residuals for the imputation model over the first 200 iterations under DA-CME NN Q=10.

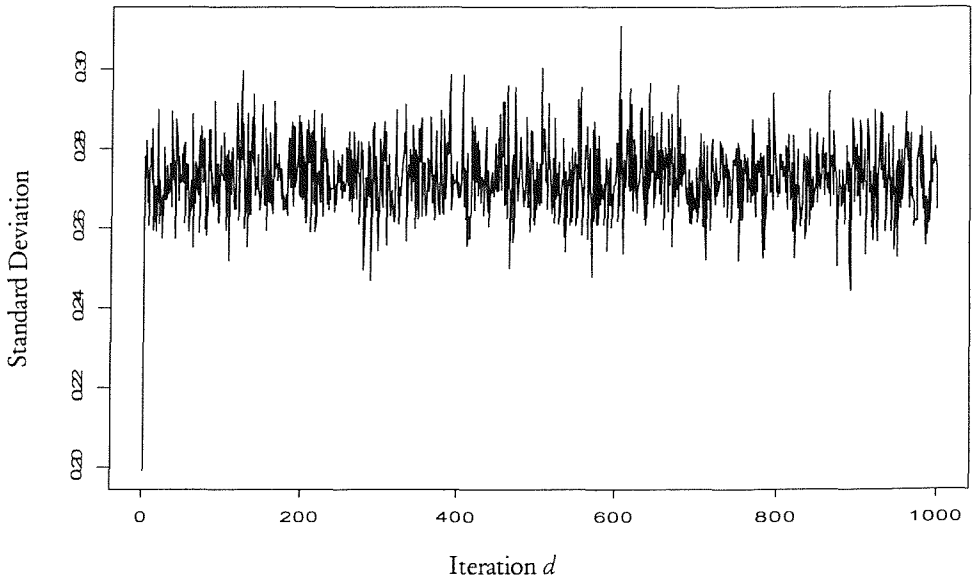

Figure A 6.6.6: Time series plot for the standard deviation of the residuals for the imputation model over the first 1000 iterations under DA-CME NN Q=10.

A 6.7) Selected Sample Autocorrelation Function Plots for Data Augmentation based on the CME Assumption using Random Regression Imputation

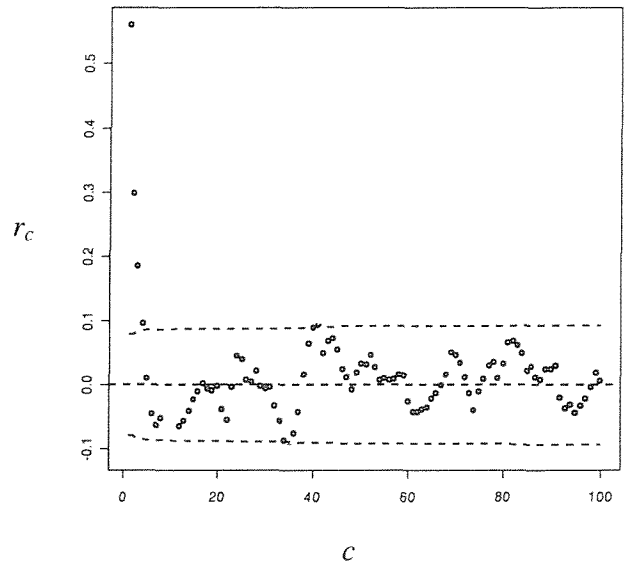

Figure A 6.7.1: Sample autocorrelation plot for the coefficient of the derived variable X in the imputation model for c=1 to 100, D=1100, under DA-CME reg imp.

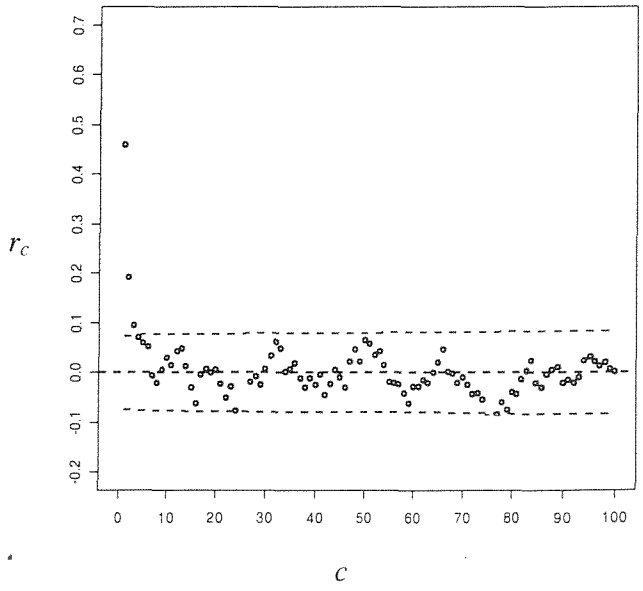

Figure A 6.7.2: Sample autocorrelation plot for the coefficient of the variable age in the imputation model for c=1 to 100, D=1100, under DA-CME reg imp.

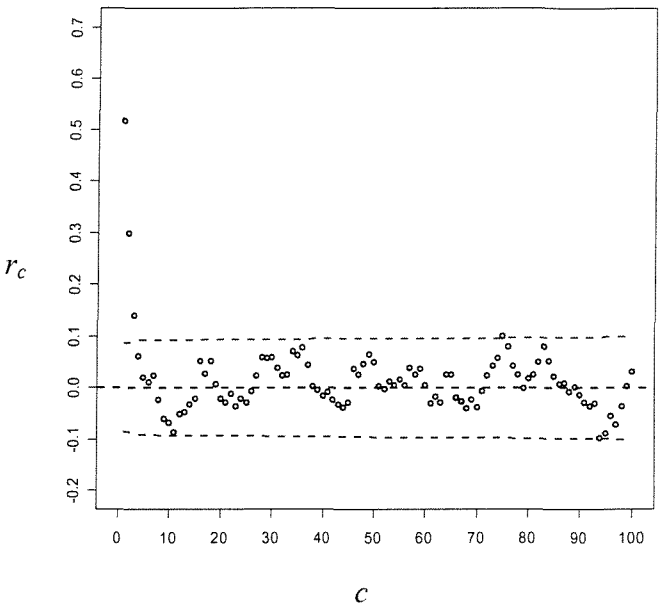

Figure A 6.7.3: Sample autocorrelation plot for the standard deviation of the residuals from the imputation model for c=1 to 100, D=1100, under DA-CME reg imp.

A 6.8) Selected Time Series Plots for Data Augmentation under CME based on the Full Nonresponse Model $f(I = 0 \mid Y, X, W)$

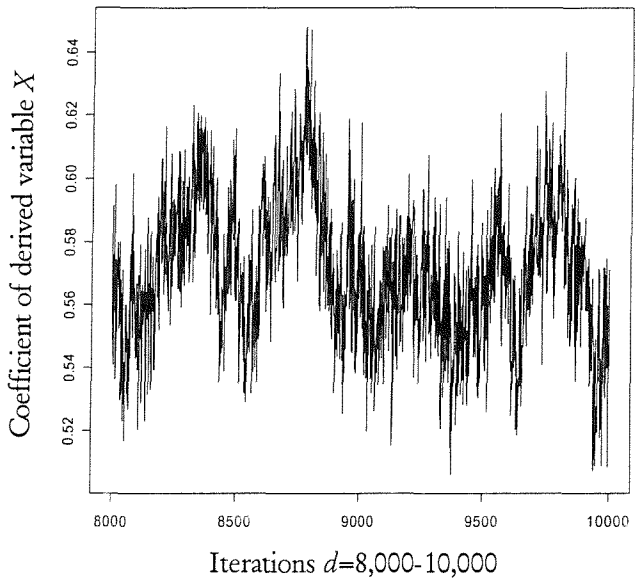

Figure A 6.8.1: Time series plot for the coefficient of the derived variable in the imputation model over iterations 8,000-10,000 under DA-CME reg imp.

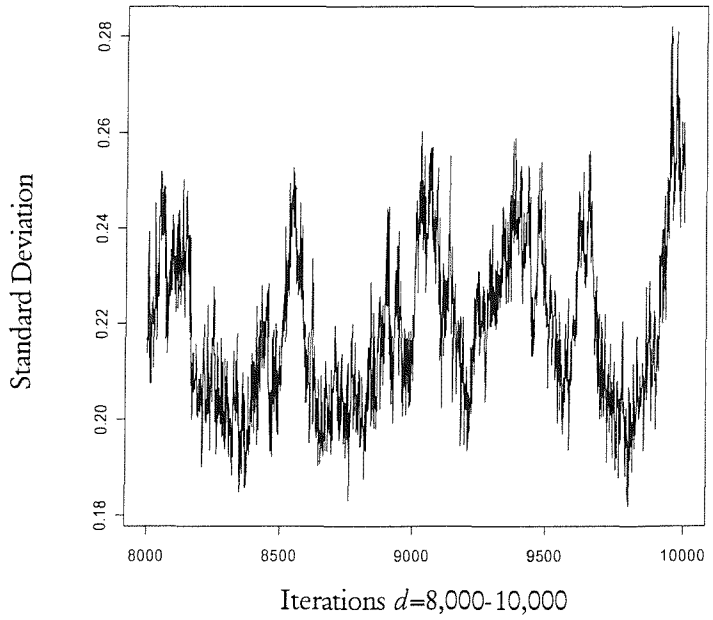

Figure A 6.8.2: Time series plot for the standard deviation of the residuals for the imputation model over iterations 8,000-10,000 under DA-CME reg imp.

A 6.9) Selected Time Series Plots for Data Augmentation based on the MAR Assumption using Random Regression Imputation

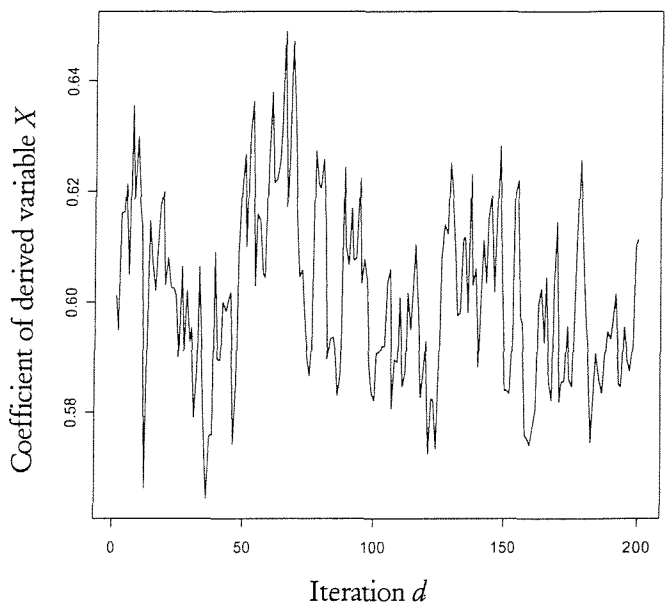

Figure A 6.9.1: Time series plot for the coefficient of the derived variable X in the imputation model over the first 200 iterations under DA-MAR reg imp.

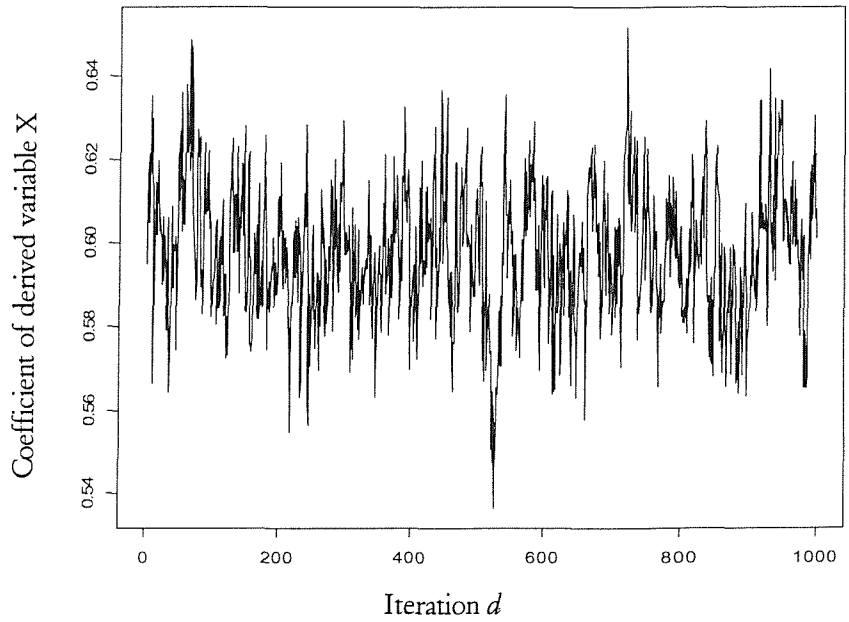

Figure A 6.9.2: Time series plot for the coefficient of the derived variable X in the imputation model over the first 1000 iterations under DA-MAR reg imp.

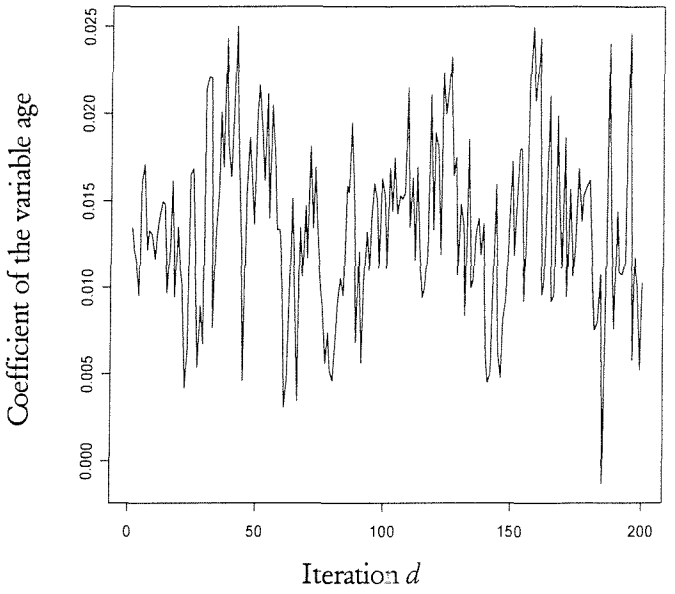

Figure A 6.9.3: Time series plot for the coefficient of the variable age in the imputation model over the first 200 iterations under DA-MAR reg imp.

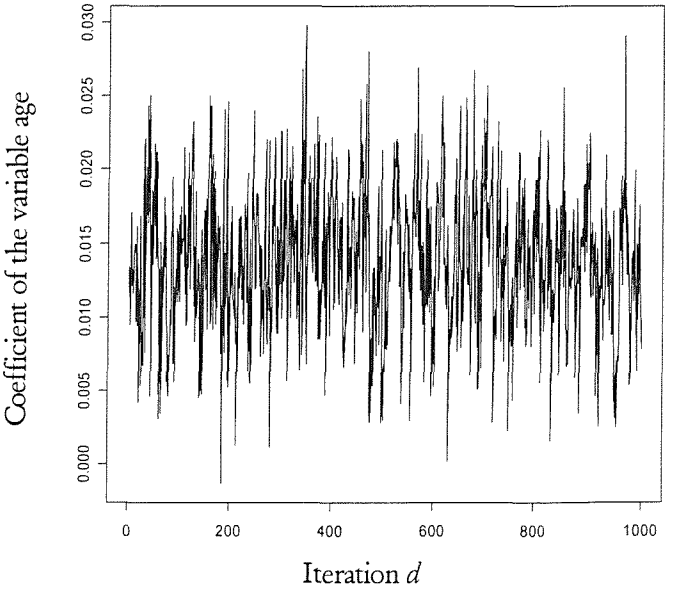

Figure A 6.9.4: Time series plot for the coefficient of the variable age in the imputation model over the first 1000 iterations under DA-MAR reg imp.

A 6.10) Selected Sample Autocorrelation Function Plots for Data Augmentation based on the MAR Assumption using Random Regression Imputation

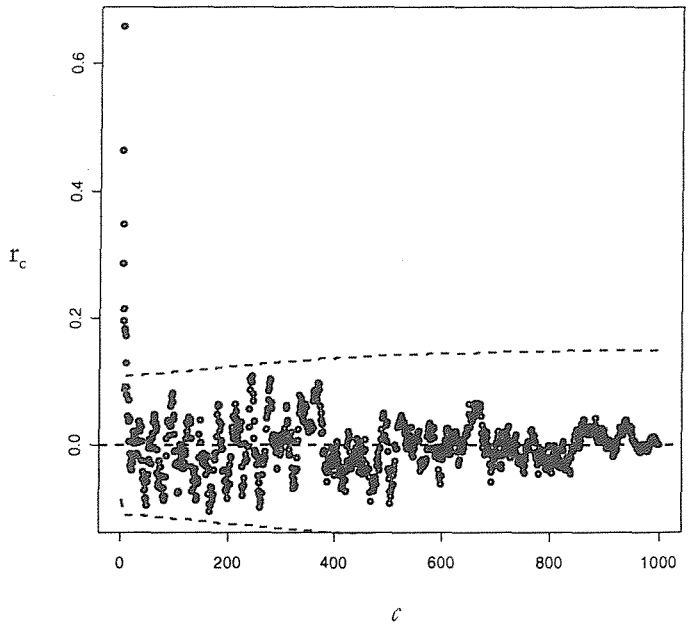

Figure A 6.10.1: Sample autocorrelation plot for the coefficient of the derived variable X in the imputation model for c=1 to 1000, D=1100, under DA-MAR reg imp.

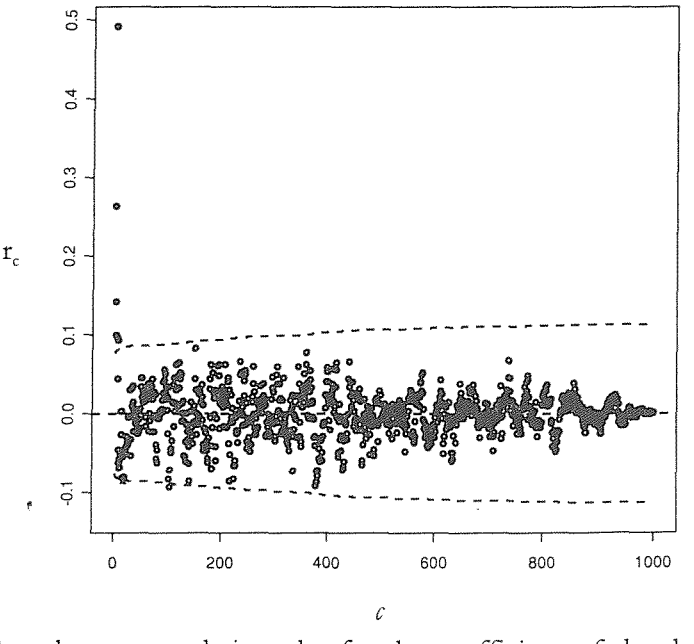

Figure A 6.10.2: Sample autocorrelation plot for the coefficient of the derived variable X in the imputation model for c=1 to 1000, D=1100, under DA-MAR reg imp.

A 6.11) Selected Time Series Plots for Data Augmentation based on the CME Assumption using Nearest Neighbour Imputation, applied to the March-May 2000 quarter.

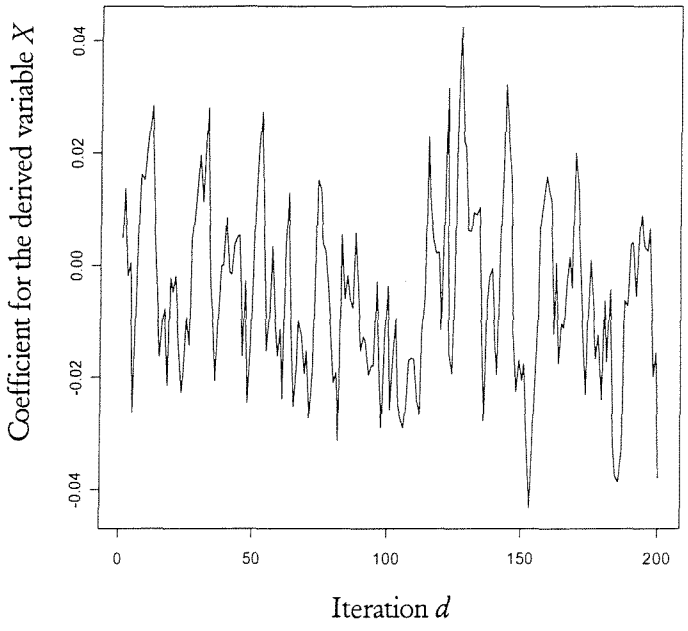

Figure A 6.11.1: Time series plot for the coefficient of the derived variable X in the imputation model over the first 200 iterations under DA-CME NN Q=10 applied to LFS.

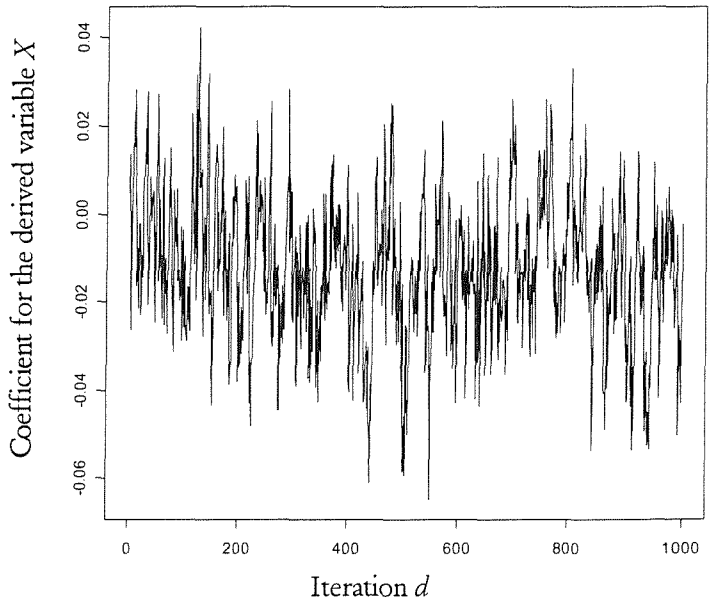

Figure A 6.11.2: Time series plot for the coefficient of the derived variable X in the imputation model over the first 1000 iterations under DA-CME NN Q=10 applied to LFS.

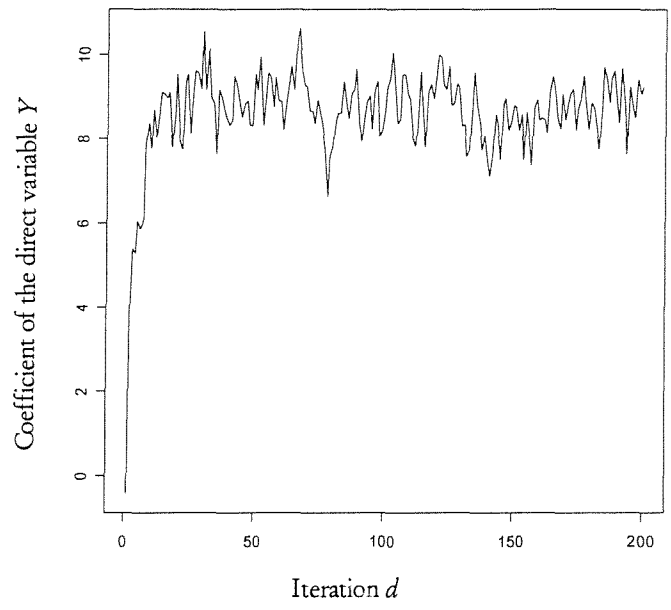

Figure A 6.11.3: Time series plot for the coefficient of the direct variable Y in the nonresponse model over the first 200 iterations under DA-CME NN Q=10 applied to LFS.

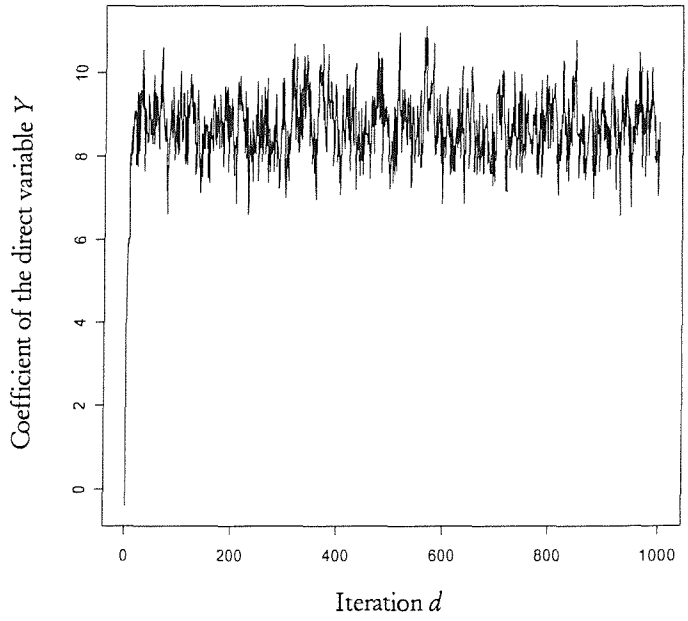

Figure A 6.11.4: Time series plot for the coefficient of the direct variable Y in the nonresponse model over the first 1000 iterations under DA-CME NN Q=10 applied to LFS.

A 6.12) Selected Sample Autocorrelation Function Plots for Data Augmentation based on the CME Assumption using Nearest Neighbour Imputation applied to the March-May 2000 quarter.

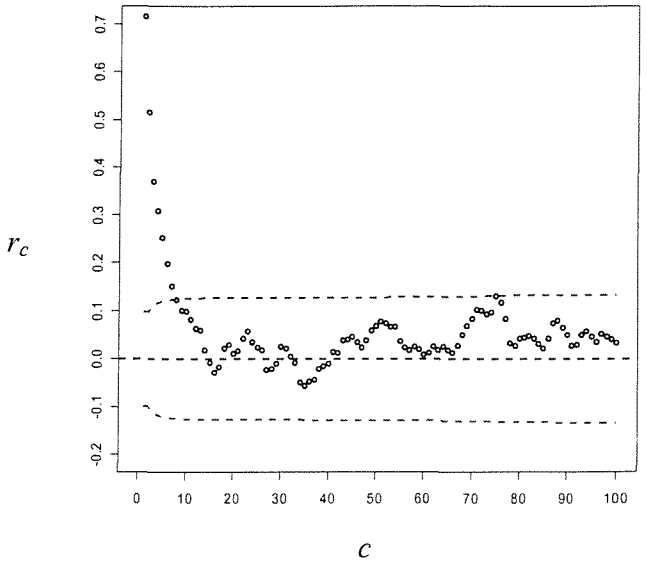

Figure A 6.12.1: Sample autocorrelation plot for the coefficient of the derived variable X in the imputation model for c=1 to 100, D=1100, under DA-CME NN Q=10.

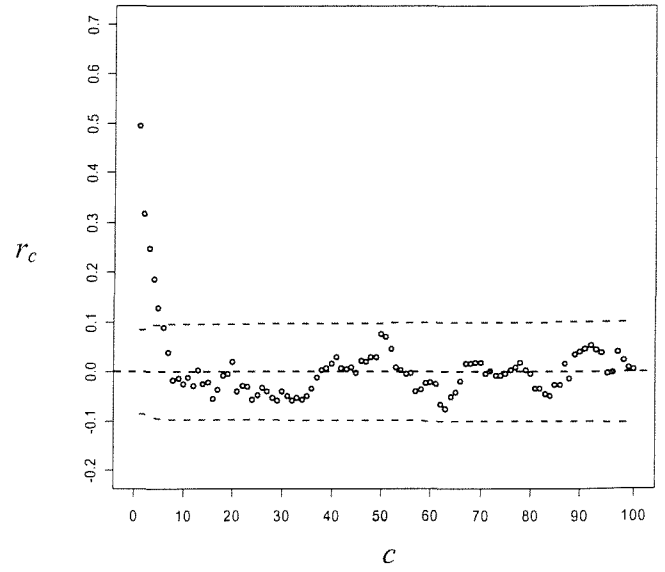

Figure A 6.12.2: Sample autocorrelation plot for the coefficient of the derived variable X in the imputation model for c=1 to 100, D=1100, under DA-CME NN Q=10.

Bibliography

[1] Agostino, R.B. and Rubin, D.B. (2000): Estimating and Using Propensity Scores with Partially Missing Data, *Journal of the American Statistical Association, Application and Case Studies*, **95**, 451, pp. 749-759.

[2] Atkinson, A.B. (1996): Seeking to Explain the Distribution of Income, in: Hills, J.: *New Inequalities*, Cambridge, pp. 19-48.

[3] Bankier, M., Fillion, J.M. and Luc, M. (1994): Imputing Numeric and Qualitative Variables Simultaneously, *Proceedings of the Section on Survey Research Methods, American Statistical Association*, pp. 242-247.

[4] Bankier, M., Luc, M. and Nadeau, C. (1995): Additional Details on Imputing Numeric and Qualitative Variables Simultaneously, *Proceedings of the Section on Survey Research Methods, American Statistical Association*, pp. 287-292.

[5] Bedrick, E. J., Christensen, R. and Johnson, W. (1996): A New Perspective on Priors for Generalised Linear Models, *Journal of the American Statistical Association, Theory and Methods*, **91**, 436, pp. 1450-1460.

[6] Bedrick, E. J., Christensen, R. and Johnson, W. (1997): Bayesian Binomial Regression: Predicting Survival at a Trauma Centre, *The American Statistician*, **51**, 3, pp. 211-218.

[7] Beissel, G. (2002): Variance Estimation for Estimates of a Pay Distribution based on Imputed Survey Data, Working paper, Department of Social Statistics, University of Southampton.

[8] Beissel-Durrant, G. and Skinner, C. (2003): Estimation of the Distribution of Hourly Pay from Household Survey Data: The Use of Missing Data Methods to Handle Measurement Error, submitted to Journal of Business and Economic Statistics, pp. 1-28.

[9] Biemer, P. and Stokes, S.L. (1991): Approaches to the Modelling of Measurement Error, in: Biemer et al. (1991), *Measurement Errors in Surveys*, pp. 487-516.

[10] Biemer, P.P. and Trewin, D. (1997): A Review of Measurement Error Effects on the Analysis of Survey Data, in: Lyberg, L. et al. (1997): *Survey Measurement and Process Quality*, New York, Chichester, pp. 603-633.

[11] Biemer, P.P., Groves, R.M., Lyberg, L., Mathiowetz, N.A. and Sudman, S. (1991): *Measurement Errors in Surveys*, New York, Chichester.

[12] Binder, D.A. and Sun, W. (1996): Frequency Valid Multiple Imputation for Surveys with a Complex Design, *Proceedings of the Section on Survey Research Methods, American Statistical Association*, pp. 281-286.

[13] Bishop,Y.M.M., Fienberg, S.E. and Holland P.W. (1978): *Discrete Multivariate Analysis, Theory and Practice*, Cambridge.

[14] 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, pp. 1-24.

[15] Bound, J., Brown, C., Duncan, G.J and Rodgers, W.L. (1990): Measurement Error in Cross-Sectional and Longitudinal Labor Market Data: Validation Study Evidence, in: Hartog, J., Ridder, G. and Theeuwes J. (eds.): *Panel Data and Labor Market Studies*, New York, pp. 1-19.

[16] 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**, 3, pp. 345-368.

[17] Box, G.E.P. and Tiao, G.G. (1992): Bayesian Inference in Statistical Analysis, Reading, 1992.

[18] Brown, P.J. (1993): *Measurement, Regression and Calibration*, Oxford, 1993.

[19] Brownstone, D. and Valetta, R.G. (1996): Modelling Earnings Measurement Error, A Multiple Imputation Approach, *Review of Economics and Statistics*, **78**, 4, pp. 705-717.

[20] Buonaccorsi, J.P. (1990): Double Sampling for Exact Values in Some Multivariate Measurement Error Problems, *Journal of the American Statistical Association*, **85**, 412, pp. 1075-1082.

[21] Carrol, R.J. and Ruppert, D. (1988): *Transformation and Weighting in Regression*, New York, London.

[22] Carrol, R.J., Ruppert, D. and Stefanski, L.A. (1995): *Measurement Error in Nonlinear Models*, London, Glasgow.

[23] Carroll, R.J and Hall, P. (1988): Optimal Rates of Convergence for Deconvolving a Density, *Journal of the American Statistical Association*, **83**, pp. 1184-186.

[24] Chen, J. and Shao, J. (1997): Biases and Variances of Survey Estimators Based on Nearest Neighbour Imputation, *Proceedings of the Section on Survey Research Methods, American Statistical Association*, pp. 365-369.

[25] Chen, J. and Shao, J. (1999): Jackknife Variance Estimation for Nearest Neighbour Imputation, *Proceedings of the Section on Survey Research Methods, American Statistical Association*, pp. 122-130.

[26] Chen, J. and Shao, J. (2000): Nearest Neighbour Imputation for Survey Data, *Journal of Official Statistics*, **16**, 2, pp. 113-131.

[27] Chen, J. and Shao, J. (2001): Jackknife Variance Estimation for Nearest Neighbour Imputation, *Journal of the American Statistical Association*, **96**, 453, pp. 260-269.

[28] Cochran, W. G. (1977): *Sampling Techniques*, New York, London.

[29] Colledge, M.J., Johnson J.H. and Pare, R. (1978): Large Scale Imputation of Survey Data, *Proceedings of the Section on Survey Research Methods, American Statistical Association*, pp. 431-435.

[30] Cowles, M.K. and Bradley, P.C. (1996): Markov Chain Monte Carlo Convergence Diagnostics: A Comparative Review, *Journal of the American Statistical Association*, **91**, 434, pp. 883-904.

[31] David, M., Little, R.J.A., Samuhel, M.E. and Triest, R.K. (1986): Alternative Methods for CPS Income Imputation, *Journal of the American Statistical Association*, **81**, 393, pp. 29-41.

[32] David, M.H., Little, R., Samuhel, M. and Triest, R. (1983): Imputation Models Based on the Propensity to Respond, *Proceedings of the Business and Economic Statistics Section, American Statistical Association*, pp. 168-173.

[33] Deville, J.C. and Särndal, C.E. (1994): Variance Estimation for the Regression Imputed Horvitz-Thompson Estimator, *Journal of Official Statistics*, **10**, 4, pp. 381-394.

[34] Dickens, R. and Manning, A. (2002): Has the National Minimum Wage Reduced UK Wage Inequality?, Paper prepared for Royal Statistical Society Conference on 'Explanations for Rising Economic Inequality', Centre for Economic Performance, London School of Economics and Political Science, June, available from:

http://cep.lse.ac.uk/pubs/download/dp0533.pdf

[35] Draper, N.R. and Smith, H. (1998): *Applied Regression Analysis*, New York, Chichester.

[36] 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**, 4, pp. 508-532.

[37] Efron, B. (1994): Missing Data, Imputation and the Bootstrap, *Journal of the American Statistical Association*, **89**, 426, pp. 463-475.

[38] Efron, B. and Tibshirani, R.J. (1993): *An Introduction to the Bootstrap*, New York, London.

[39] Ernst, L.R. (1980): Variance of the Estimated Mean for Several Imputation Procedures, *Proceedings of the Section on Survey Research Methods, American Statistical Association*, pp. 716-720.

[40] Fay, R.E. (1996): Alternative Paradigms for the Analysis of Imputed Survey Data, *Journal of the American Statistical Association*, **91**, 434, pp. 490-498.

[41] Fay, R.E. (1999): Theory and Application of Nearest Neighbour Imputation in Census 2000, *Proceedings of the Section on Survey Research Methods, American Statistical Association*, pp. 112-121.

[42] Fuller, W.A. (1987): *Measurement Error Models*, New York, Chichester.

[43] Fuller, W.A. (1995): Estimation in the Presence of Measurement Error, *International Statistical Review*, **63**, 2, pp. 121-147.

[44] Fuller, W.A. and Guenther, P.M. (1997): Estimating Usual Dietary Intake Distributions, in: Lyberg, L. et al. (1997): *Survey Measurement and Process Quality*, New York, Chichester, pp. 689-709.

[45] Fuller, W.A. and Kim, J.K. (2002): Hot Deck Imputation for the Response Model, working paper.

[46] Gagnon, F., Lee, H., Provost, M., Rancourt, E. and Särndal, C.E. (1997): Estimation of Variance in Presence of Imputation, *Proceedings of Statistics Canada Symposium 97, New Directions in Surveys and Censuses*, pp. 30-42.

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

[48] Gilks, W.R., Richardson, S. and Spiegelhalter, D.J. (1996): *Markov Chain Monte Carlo in Practice*, London.

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

[50] Greenlees, J.S., Reece, W.S., Zieschang, K.D. (1982): Imputation of Missing Values when the Probability of Response Depends on the Variable Being Imputed, *Journal of the American Statistical Association*, 77, 378, pp. 251-261.

[51] Griffiths, A. and Wall, S. (1999): *Applied Economics*, London, New York.

[52] Hampel, F.R, Ronchetti, E.M., Rousseuw, P.J. and Stahel, W. (1986): *Robust Statistics*, New York.

[53] Heitjan, D.F. (1994): Ignorability in General Incomplete-Data Models, *Biometrika*, **81**, 4, pp. 701-708.

[54] Heitjan, D.F. and Landis, J.R. (1994): Assessing Secular Trends in Blood Pressure, A Multiple Imputation Approach, *Journal of the American Statistical Association*, **89**, 427, pp. 750-759.

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

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

[57] Holmes, D.J and Skinner, C.J. (2000): Variance Estimation for Labour Force Survey Estimates of Level and Change, *Government Statistical Service, Methodology Series No. 21*, London.

[58] Jaech, J.L. (1985): *Statistical Analysis of Measurement Error*, New York, Chichester.

[59] Jones, S.M. and Chromy, J.R. (1982): Improved Variance Estimators Using Weighting Class Adjustments for Sample Survey Nonresponse, *Proceedings of the Survey Research Methods Section, American Statistical Association*, pp. 105-110.

[60] Kalton, G. (1983): *Compensating for Missing Survey Data*, Michigan.

[61] Kalton, G. and Kasprzyk, D. (1982): Imputing for Missing Survey Responses, *Proceedings of the Section on Survey Research Methods, American Statistical Association*, pp. 22-31.

[62] Kalton, G. and Kasprzyk, D. (1986): The Treatment of Missing Survey Data, *Survey Methodology*, **12**, 1, pp. 1-16.

[63] Kalton, G. and Kish, L. (1984): Some Efficient Random Imputation Methods, *Communications in Statistics, Part A, Theory and Methods*, 13, pp. 1919-1939.

[64] Kim, J.K. (2002): A Note on Approximate Bayesian Bootstrap Imputation, *Biometrika*, **89**, 2, pp. 470-477.

[65] Kott, P.S. (1994): A Note on Handling Nonresponse in Sample Surveys, *Journal of the American Statistical Association*, **89**, 426, pp. 693-696.

[66] Kovar, J.G. and Chen, E.J. (1994): Jackknife Variance Estimation of Imputed Survey Data, *Survey Methodology*, **20**, 1, pp. 45-52.

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

[68] Kuha, J. and Skinner, C. (1997): Categorical Data Analysis and Misclassification, in: Lyberg, L. et al. (1997): *Survey Measurement and Process Quality*, New York, Chichester, pp. 633-670.

[69] Kuk, A.Y.C., Mak, T.K. and Li, W.K. (2001): Estimation Procedures for Categorical Survey Data with Nonignorable Nonresponse, *Communications in Statistics, Theory and Methods*, **30**, 4, pp. 643-663.

[70] Laaksonen, S.S. (1991): Adjustment for Non-Response in Two-Year Panel Data, Applications to Problems of Household Income Distribution, *Statistician, Special Issue: Survey Design, Methodology and Analysis*, **40**, 2, pp. 153-168.

[71] Landerman, L.R., Land, K.C. and Pieper, C.F (1997): An Empirical Evaluation of the Predictive Mean Matching Method for Imputing Missing Values, *Sociological Methods and Research*, **26**,1, pp. 3-33.

[72] Lee, H. and Rancourt, E. (1991): Experiments with Variance Estimation from Survey Data with Imputed Values, *Proceedings of the Survey Research Methods Section, American Statistical Association*, pp. 690-695.

[73] Lee, H., Rancourt, E. and Särndal, C.E. (1994): Experiments with Variance Estimation from Survey Data with Imputed Values, *Journal of Official Statistics*, **10**, 3, pp. 231-243.

[74] Lee, H., Rancourt, E. and Särndal, C.E. (1995): Variance Estimation in the Presence of Imputed Data for the Generalized Estimation System, *Proceedings of the Section on Survey Research Methods, American Statistical Association*, pp. 384-389.

[75] Lee, H., Rancourt, E. and Särndal, C.E. (2000): Variance Estimation from Survey Data under Single Value Imputation, Working Paper, Methodology Branch, Household Survey Methods Division Statistics Canada, pp. 1-30.

[76] Lee, H., Rancourt, E. and Särndal, C.E. (2002): Variance Estimation from Survey Data under Single Value Imputation, in: Groves, R.M., Dilman, D.A., Eltinge, J.L. and Little, R.J.A. (2002): *Survey Nonresponse*, New York, pp. 315-328.

[77] Leonard, T. (1972): Bayesian Methods for Binomial Data, *Biometrika*, **59**, 3, pp. 581-588.

[78] Lepkowski, J.M., Stehouwer, S.A. and Landis, J.R. (1984): Strategies for the Analysis of Imputed Data in a Sample Survey, *Proceedings of the Section on Survey Research Methods, American Statistical Association*, pp. 622-627.

[79] Lessler, J.T. (1983): An Expanded Survey Error Model, in: Madow, W.G. and Olkin, J. (eds.), *Incomplete Data in Sample Surveys, Proceedings of the Symposium*, 3, New York, London, pp. 259-270.

[80] Lessler, J.T. and Kalsbeek W.D. (1992): *Nonsampling Error in Surveys*, New York, Chichester.

[81] Lipsitz, S.R., Zhao, L.P. and Molenberghs, G. (1998): A Semiparametric Method of Multiple Imputation, *Journal of the Royal Statistical Society, Series B, Statistical Methodology*, **60**, 1, pp. 127-144.

[82] Little, R.J.A. (1982): Models for Nonresponse in Sample Surveys, *Journal of the American Statistical Association*, 77, 378, pp. 237-250.

[83] Little, R.J.A. (1986): Survey Nonresponse Adjustments for Estimates of Means, *International Statistical Review*, **54**, 2, pp. 139-157.

[84] Little, R.J.A. (1988, Dec.): A Test of Missing Completely at Random for Multivariate Data with Missing Values, *Journal of the American Statistical Association*, **83**, 404, pp. 1198-1202.

[85] Little, R.J.A. (1988, July): Missing-Data Adjustments in Large Surveys, *Journal of Business and Economic Statistics*, **6**, 3, pp. 287-301.

[86] Little, R.J.A. and Rubin, D.B. (2002): *Statistical Analysis with Missing Data*, New York, Chichester.

[87] Low Pay Commission (1998): *The National Minimum Wage*, First Report of the Low Pay Commission, London.

[88] Low Pay Commission (2000): *The National Minimum Wage*, Second Report of the Low Pay Commission, London.

[89] Low Pay Commission (2001): *The National Minimum Wage*, Third Report of the Low Pay Commission, vol. 1 and 2, London.

[90] Low Pay Commission (2003): *The National Minimum Wage*, Fourth Report of the Low Pay Commission, London.

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

[92] Lyberg, L., Biemer, P., Collins, M., Leeuw, de E., Dippo, C., Schwarz, N. and Trewin, D. (1997): *Survey Measurement and Process Quality*, New York, Chichester.

[93] Manning, A. and Dickens, R. (2002): The Impact of the National Minimum Wage on the Wage Distribution, Poverty and the Gender Pay Gap, working paper prepared for the Low Pay Commission, pp. 1-98.

[94] Mardia, K.V., Kent, J.T. and Bibby, J.M. (1979): *Multivariate Analysis*, London, 1979.

[95] Moore, J.C., Stinson, L.L. and Welniak, J.E. (2000): Income Measurement Error in Surveys, A Review, *Journal of Official Statistics*, **16**, 4, pp. 331-361.

[96] Moustaki, I. and Knott, M. (2000): Weighting for Item Non-Response in Attitude Scales by using Latent Variable Models with Covariates, *Journal of the Royal Statistical Society, Series A*, **163**, 3, pp. 445-459.

[97] Neuilly, M. and Cetama, B. (1999): *Modelling and Estimation of Measurement Errors*, London, Paris, New York.

[98] Nordberg, L., Penttilae, I. and Sandstroem, S. (2001): Earnings Data from Surveys and Registers, paper presented at the CHINTEX Workshop, Nov. 2001, pp. 1-12.

[99] Nusser, S.M., Carriquiry, A.L., Dodd, K.W. and Fuller, W.A. (1996): A Semiparametric Transformation Approach to Estimating Usual Daily Intake Distributions, *Journal of the American Statistical Association, Applications and Case Studies*, **91**, 463, pp. 1440-1449.

[100] Office for National Statistics (ONS), Government Statistical Service (1996): Report of the Task Force on Imputation, GSS Methodology Series no. 3, London.

[101] Oh, H.L. and Scheuren, F.J. (1980): Estimating the Variance Impact of Missing CPS Income Data, *Proceedings of the Section on Survey Research Methods, American Statistical Association*, pp. 131-138.

[102] Oh, H.L. and Scheuren, F.J. (1983): Weighting Adjustment for Unit Nonresponse, in: Madow, W.G., Olkin, I. and Rubin, D.B. (eds.), *Incomplete Data in Sample Surveys*, Vol. 2, Theory and Bibliographies, New York, pp. 143-183.

[103] ONS (1999): Labour Force Survey, User Guide, Vol. I, Background and Methodology, London.

[104] ONS (1999, Oct.): New LFS Variable HRRATE, What is Your Hourly Rate?, unpublished report.

[105] ONS (2000): Improving Low Pay Estimates, unpublished report.

[106] Pischke, J.S. (1995): Measurement Error and Earnings Dynamics, Some Estimates From PSID Validation Study, *Journal of Business and Economic Statistics*, **13**, 3, pp. 305-314.

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

[108] Rancourt, E. (1999): Estimation with Nearest Neighbour Imputation at Statistics Canada, in: *Proceedings of the Section on Survey Research Methods, American Statistical Association*, pp. 131-138.

[109] Rancourt, E., Särndal, C. and Lee, H. (1994): Estimation of the Variance in the Presence of Nearest Neighbour Imputation, *Proceedings of the Survey Research Methods Section, American Statistical Association*, pp. 888-893.

[110] Rao, J.N.K. (1996): On Variance Estimation with Imputed Survey Data, *Journal of the American Statistical Association*, **91**, 434, pp. 499-506.

[111] Rao, J.N.K. (2001): Variance Estimation in the Presence of Imputation for Missing Data, unpublished paper, pp. 1-12.

[112] Rao, J.N.K. and Shao, J. (1992): Jackknife Variance Estimation with Survey Data under Hot Deck Imputation, *Biometrika*, **79**, 4, pp. 811-822.

[113] Rao, J.N.K. and Sitter, R.R. (1995): Variance Estimation under Two-Phase Sampling with Applications to Imputation for Missing Data, *Biometrika*, **82**, 2, pp. 453-460.

[114] Rao, J.N.K. and Sitter, R.R. (1997): Variance Estimation under Stratified Two-Phase Sampling with Applications to Measurement Bias, in: Lyberg, L. et al. (1997): *Survey Measurement and Process Quality*, New York, Chichester, pp. 753-768.

[115] Reilly, M. (1993): Data Analysis using Hot Deck Multiple Imputation, *Statistician, Special Issue: Conference on Applied Statistics in Ireland*, **42**, 3, pp. 307-313.

[116] Rodgers, W.L. and Herzog, R. (1987): Covariances of Measurement Errors in Survey Responses, *Journal of Official Statistics*, **3**, 4, pp.403-418.

[117] 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**, 424, pp. 1208-1218.

[118] Rosenbaum and Rubin (1983): The Central Role of the Propensity Score in Observational Studies for Causal Effects, *Biometrika*, 70, 1, pp. 41-55.

[119] Rosenthal, J.S. (1993): Rates of Convergence for Data Augmentation on Finite Sample Spaces, *The Annals of Applied Probability*, **3**, 3, pp. 819-839.

[120] Rubin, D.B. (1976): Inference and Missing Data, *Biometrika*, **63**, pp. 581-592.

[121] Rubin, D.B. (1987): *Multiple Imputation for Nonresponse in Surveys*, New York, Chichester.

[122] Rubin, D.B. (1994): Missing Data, Imputation and the Bootstrap: Comment, *Journal of the American Statistical Association*, **89**, 426, pp. 475-478.

[123] Rubin, D.B. (1996): Multiple Imputation after 18+ Years, *Journal of the American Statistical Association*, **91**, 434, pp. 473-489.

[124] Rubin, D.B. and Schenker. N. (1986): Multiple Imputation for Interval Estimation from Simple Random Samples With Ignorable Nonresponse, *Journal of the American Statistical Association*, **81**, 394, pp. 366-374.

[125] Sande, I.G. (1982): Imputation in Surveys: Coping with Reality, *The American Statistician*, **36**, 3, pp. 145-152.

[126] Sande, I.G. (1988): Comment to Little, R.A. 1988, *Journal of Business and Economic Statistics*, **6**, 3, pp. 296-297.

[127] Saris, W.E. and Andrews F.M. (1991): Evaluation of Measurement Instruments Using a Structural Modelling Approach, in: Biemer et al. (1991), *Measurement Errors in Surveys*, pp. 575-598.

[128] Särndal, C.E., Swensson, B. and Wretman, J. (1992): *Model Assisted Survey Sampling*, New York.

[129] Schafer, J.L. (1997): *Analysis of Incomplete Multivariate Data*, London.

[130] Schenker, N. and Taylor, J.M.G. (1996): Partially Parametric Techniques for Multiple Imputation, *Computational Statistics and Data Analysis*, **22**, pp. 425-446.

[131] Scheuren, F. (1988): Comment to Little, R.A., 1988, *Journal of Business and Economic Statistics*, **6**, 3, pp. 298-299.

[132] Schulte Nordholt, E.S. (1998): Imputation: Methods, Simulation, Experiments and Practical Examples, *International Statistical Review*, **66**, 2, pp. 157-180.

[133] Selén, J. (1986): Adjusting for Errors in Classification and Measurement in the Analysis of Partly and Purely Categorical Data, *Journal of the American Statistical Association*, **81**, 393, p. 75-80.

[134] Shao, J and Sitter, R.R. (1996): Bootstrap for Imputed Survey Data, *Journal of the American Statistical Association, Theory and Methods*, **91**, 435, pp. 1278-1286.

[135] Shao, J. and Steel, P. (1999): Variance Estimation for Survey Data with Composite Imputation and Nonnegligible Sampling Fractions, *Journal of the American Statistical Association, Theory and Methods*, **94**, 445, pp. 254-265.

[136] Shao, J. and Wu, C.F.J. (1989): A General Theory for Jackknife Variance Estimation, *The Annals of Statistics*, **17**, 3, pp. 1176-1197.

[137] Shao, J., Chen, Y. and Chen Y. (1998): Balanced Repeated Replication for Stratified Multistage Survey Data Under Imputation, *Journal of the American Statistical Association, Theory and Methods*, **93**, 442, pp. 819-831.

[138] Sitter, R.R. and Rao, J.N.K. (1997): Imputation for Missing Values and Corresponding Variance Estimation, *The Canadian Journal of Statistics*, **25**, 1, pp. 61-73.

[139] Skinner, C. (1994): Measurement Errors and Survey Data Analysis, *The Survey Statistician*, **30**, pp. 15-19.

[140] Skinner, C. and Beissel, G. (2001): Estimating the Distribution of Hourly Pay from Survey Data, paper presented at the CHINTEX Workshop, The Future of Social Surveys in Europe, Helsinki 29, 30 Nov. 2001, pp. 1-15.

[141] Skinner, C. and Rao, J.N.K. (2002): Jackknife Variance Estimation for Multivariate Statistics under Hot Deck Imputation From Common Donors, *Journal of Statistical Planning and Inference*, **102**, 1, pp. 421-422.

[142] 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**, pp. 653-676.

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

[144] Stefanski, L.A. and Carroll, R.J. (1987): Conditional Scores and Optimal Scores in Generalised Linear Measurement Error Models, *Biometrika*, **74**, 703-716.

[145] Stefanski, L.A. and Carroll, R.J. (1990): Deconvoluting Kernel Density Estimators, *Statistics*, **21**, 165-184.

[146] Stuttard, N. and Jenkins, J. (2001): Measuring Low Pay Using the New Earnings Survey and the Labour Force Survey, *Labour Market Trends*, pp. 55-66.

[147] Sunley, P. and Martin, R. (2000): The Geographies of the National Minimum Wage, *Environment and Planning A*, **32**, pp. 1735-1758.

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

[149] Tanner, M.A. and Wong, W.H. (1987): The Calculation of Posterior Distributions by Data Augmentation, *Journal of the American Statistical Association*, **82**, 398, pp. 528-540.

[150] Tollefson, M. and Fuller, W.A. (1992): Variance Estimation for Samples with Random Imputation, *Proceedings of the Section on Survey Research Methods*, *American Statistical Association*, pp. 758-763.

[151] Tsutakawa, R.K. and Lin, H.Y. (1986): Bayesian Estimation of Item Response Curves, *Psychometrika*, **51**, 2, pp. 251-267.

[152] Whittaker, J. (1990): *Graphical Models in Applied Multivariate Statistics*, Chichester, New York.

[153] Wilkinson, D. (1998, Dec.): Who are the Low Paid, *Labour Market Trends*, pp. 617-622.

[154] Wilkinson, D. (1998, May): Towards Reconciliation of NES and LFS Earnings Data, *Labour Market Trends*, pp. 223-229.

[155] Wolter, K.M. (1985): *Introduction to Variance Estimation*, New York, Berlin.

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