Missing Observations in Survey Data: An Experimental Approach to Imputation

Melvyn Weeks
Faculty of Economics and Politics
University of Cambridge,
Sidgwick Avenue, Cambridge CB3 9DE

Alan Hughes
Centre for Business Research
University of Cambridge
Sidgwick Avenue, Cambridge CB3 9DE

PRELIMINARY FIRST DRAFT
DO NOT QUOTE

April 3, 2001

Abstract

The collection and analysis of large micro datasets is now a critical input into policy making, both in terms of analysing and understanding the decision of consumers and firms. However, despite an increased awareness of the importance of survey design, the existence of both item and questionnaire non-response represents a significant problem both in terms of the information lost, and potential biases in the presence of non-random non-response. Given the availability of a secondary data source for non-respondents, we utilise an experimental approach to evaluate the performance over a range of imputation techniques.

`http://www.econ.cam.ac.uk/faculty/weeks/MissExp/MissExp.pdf`

Acknowledgements: The authors acknowledge financial support of the Economic and Social Research Council, award number R000222824.

JEL Classification: C10, C24, C81, C88

Key Words: Missing data; non-random non-responses; imputation methods; business surveys.

1. Introduction

> *Survey nonresponse is problematic for identification of population parameters. Whether nonresponse takes the form of particular missing items or entire missing interviews, the only way to identify population parameters is to make assumptions that determine the distribution of the missing data. A basic problem of empirical analysis is that such assumptions are not testable.*
>
> Horowitz and Manski (1998)

Based upon recent experience with national censuses taken both in the US and UK, over a quarter of all responses will contain some form of item non-response. For example, Lillard, Smith, and Welch (1986) note that non-response to the census income question has increased from 2.5% in 1940 to 26% in 1982. On a slightly less grand scale a large proportion of both ESRC and NSF funded projects involve the creation of complex datasets, many of which are based upon sample survey of firms or individuals. Since these datasets represent an important source of information for policy makers, the significant costs of data collection has resulted in considerable effort to ensure both the accuracy of the information and examine the impact of unit attrition and item non-response. In this context the problem of missing data and the related issues such as selection bias is paramount.

At the outset we emphasise the difference between imputation of missing data and weighting the *observed* data to provide appropriate adjustments for biases due to non-random selection. This follows naturally from the distinction between *end-users* of a database and the *providers*. In econometrics, and following the work of Heckman (1979), end-users have developed a wealth of techniques, both parametric and non-parametric, to correct for the effects of non-random nonresponse.[1] However, these corrections are, in general, user specific and therefore reflect the considerable heterogeneity across users in both the understanding of the consequence of non-response, and the ability to correct for potential biases. Aligned with the increasing cost of collecting survey data, there is pressure to

[1] This problem is often referred to as self-selection.

provide, where possible, complete datasets with imputation carried out by the database provider.

We adopt a quasi experimental approach to examine the performance of imputation techniques. We focus on the problem of item nonresponse in the Centre for Business Research database on small and medium-sized enterprises (SME's) and in particular the problem of missing profit and employment data. Analogous to the problem of evaluating non-experimental methods for controlling for sample selection, we adopt a methodology that exploits the availability of a comparable profit and employment series (the Inter Company Comparison database (ICC)[2] through the data which companies must deposit at Companies House. Notwithstanding the use of modified accounts by some small companies which do not require profits to be reported[3], this data series has complete entries for a high percentage of the CBR data covering both respondents and non-respondents. In this context we may determine the validity of both model assumptions and the specification of the model of nonresponse, thereby addressing, in part, the comments of Horowitz and Manski (1998). In a comparable study examining income nonresponse in the Current Population Survey, Greenlees, Reece, and Zieschang (1982) found that the process determining nonresponse is dependent upon the value of income. Preliminary analysis suggests that the process of nonresponse to profit is also endogenous.

2. Missing Data as Survey Non-response

In the social sciences model-based inference depends upon the use of sample observations to provide both predictive information for the phenomenon under study, and a guide to the importance of structural parameters. In this regard a critical component of a model-builders toolkit is methodologies to assist in the selection of a model for the *observed* data. If we now add an additional problem in the guise of *missing* data, a number of additional issues must be faced. Critical in this respect is whether the process which generates the missing data can be ignored,

[2]The ICC (Inter Company Comparisons) database is the largest on-line database covering the whole of the UK company sector. It forms a component of both the Datastream and OneSource databases which have been used extensively in recent work on the UK company sector.

[3]We elaborate on this issue in Section 7.2.

such that valid inference can proceed using only the observed data.

Even in the case of data which is missing at random where model parameters are, in general, unbiased and consistent, there will be an efficiency loss thereby compromising the validity of model estimates through wider confidence bands. In instances where the analyst recognises that the missing data problem is one of stochastic censoring and subsequently incorporates a probability of response model, the major drawback is the need to make non-testable assumptions about the distribution of the missing data. Weeks (2001) has provided an extensive overview of the missing data problem including a taxonomy of a wide range of conditional and unconditional imputation techniques.

When observational rules are generated by non-random processes which are intrinsically related to the phenomenon of interest, the resulting observed data does not represent an ignorable random sample of the underlying data. In the case of non-response in survey data we may consider a number of observational rules (or self-selection mechanisms) based upon, for example, processes that determine the likelihood of returning a questionnaire and/or the probability of non-response on a particular question. In the context of a business survey, the key issue is whether a firm chooses to respond to a particular question, and whether we can identify one or more firm characteristics which differ significantly across the respondent and non-respondent states such that we might be able to construct a model which approximates the missing data generating process (*mdgp*).

A recent study by Horowitz and Manski (1998) notes that the major problem of methods used by econometricians and statisticians to control for the problem of bias introduced by the use of non-random samples, is that unless untestable assumptions about the distribution of the missing data are made, identification of population parameters are impossible. Little (1982) also warns against the pitfalls of model-based techniques for correcting selection bias, and suggests the use of a variety of models to determine sensitivity to nonresponse bias. However, despite the recent emergence of a number of alternatives such as nonparametric techniques (see Ahn and Powell (1993)) and entropy-based methods (see Golan, Judge, and Miller (1996)), the parametric modelling approach to handle both unit and item non-response in sample surveys is still predominant. Obviously if

the missing data were available then these assumptions could be verified, together with a comparison of imputation methods which both ignored and incorporated a model of non-response in the imputation procedure.

An experimental approach to handling different types of missing data has been advocated in a number of different situations. For example, in the context of the evaluation of various manpower training programmes, the *missing data* is the performance of either the firm or an individual in the counterfactual state i.e. it is not possible to observe the same economic agent in both a treatment and non-treatment state. Policy evaluation based on random assignment of program applicants to a recipient and non-recipient (control) group has been adopted by Heckman and Smith (1995). This procedure offers an alternative solution to the problem of selection bias in that a random sample is taken from the *treatment* group which are then denied access to the program. Heckman and Hotz (1989) follow a similar approach in the evaluation of alternative non-experimental methods for estimating the impact of manpower training. Based upon a study by LaLonde (1986), the authors use both experimental and non-experimental data to evaluate non-experimental estimates of the impact of the programme. A recent study by Hurd, McFadden, Chand, Gan, Merrill, and Roberts (1997) has utilised an experimental approach to evaluate survey response bias in a dataset documenting the consumption and savings of the elderly.

3. Missing Data Processes

Below we introduce notation and define a number of missing data processes by focussing upon a data matrix $\mathbf{M} = \{m_{ij}\}$, $i = 1, ...n$, $j = 1, ..., k$, where i indexes individual firms and j covariates. In the analysis that follows, missing data will refer to a situation where data is missing for one or more rows and columns of $\mathbf{M}$. The applicability of any particular form of imputation depends, in part, upon the dimensionality of $\mathbf{M}$, whether inference is unconditional or conditional, and the pattern of missing data.[4]

We partition the n observations into two mutually exclusive sets: $\mathbf{M}_\ell$ denotes

[4]The most appropriate form of imputation will also depend upon whether univariate or multivariate analysis is to be performed.

fully observed data and $\mathbf{M}_{n-\ell}$ denotes a matrix where at least one column entry per row is missing. Note that depending upon the pattern of the missing data $\mathbf{M}_\ell$ and $\mathbf{M}_{n-\ell}$ may take various forms. For example, in the case of questionnaire non-response for a subset of firms, $\mathbf{M}_\ell$ will comprise a *rectangular* dataset with no missing values across the k columns.

Further we let $\mathbf{Q}$ denote an $n \times k$ matrix with typical element $q_{ij} = 1$ if m_{ij} is observed and zero otherwise. The extent to which data is both missing at random *(MAR)* and observed at random *(OAR)* may be analysed by examining the *joint* distribution of $\mathbf{M}$ and $\mathbf{Q}$. For example, the joint distribution of $\mathbf{M}$ and $\mathbf{Q}$ is given by

$$f(\mathbf{M}, \mathbf{Q} \mid \boldsymbol{\theta}, \boldsymbol{\beta}) = f(\mathbf{M} \mid \boldsymbol{\theta}) f(\mathbf{Q} | \mathbf{M}, \boldsymbol{\beta}), \tag{3.1}$$

where $\boldsymbol{\theta}$ and $\boldsymbol{\beta}$ are vectors of parameters, and $f(\mathbf{Q} \mid \mathbf{M}, \boldsymbol{\beta})$ is the distribution of the missing data mechanism. We will also refer to $f(.)$ as the predictive (or posterior) distribution of the missing data. $\boldsymbol{\theta}$ represents parameters of interest and $\boldsymbol{\beta}$ are parameters which determine the *mdgp*. Since the observed information consists of $\{\mathbf{M}_l, \mathbf{Q}\}$ the key issue is characterising situations when estimation and inference can be based upon $\mathbf{M}_\ell$, thus ignoring the missing data mechanism.[5] If this can be done then instead of (3.1) we can write

$$f(\mathbf{M}_\ell, \mathbf{Q} \mid \boldsymbol{\theta}, \boldsymbol{\beta}) = f(\mathbf{M}_\ell \mid \boldsymbol{\theta}) f(\mathbf{Q} \mid \mathbf{M}_\ell, \boldsymbol{\beta}).$$

It therefore follows that if we can simplify $f(\mathbf{Q} \mid \mathbf{M}, \boldsymbol{\beta})$ such that

$$f(\mathbf{Q} \mid \mathbf{M}_\ell, \mathbf{M}_{n-\ell}, \boldsymbol{\beta}) = f(\mathbf{Q} \mid \mathbf{M}_\ell, \boldsymbol{\beta}), \tag{3.2}$$

then data is missing at random or *mar.* Putting (3.2) into words we now define *mar.*

Definition 3.1. *mar*

The missing data are missing at random (mar) if the predictive distribution of the missing data, (namely $f(\mathbf{Q} \mid \mathbf{M}_\ell, \mathbf{M}_{n-\ell}, \boldsymbol{\beta})$), is the same for all possible values of the missing data such that $f(\mathbf{Q} \mid \mathbf{M}_\ell, \mathbf{M}_{n-\ell}, \boldsymbol{\beta}) = f(\mathbf{Q} \mid \mathbf{M}_\ell, \boldsymbol{\beta})$.

[5] Note that in this instance we assume that the rows of $\mathbf{M}$ are interchangeable.

In more general terms, the probability of missing data depends upon the observed but not the unobserved data. For example, if in a survey of firms there is item non response for profit data, denoted here as the $n \times 1$ vector $\mathbf{m}^p$, and that the extent of non-response is high for both small and large firms and independent of the level of profits, then the pattern of missing data is predictable using firm size. Note that if a particular missing data process is considered mar and the vector of parameters, $\boldsymbol{\beta}$, which determine the pattern of missing data ($\mathbf{Q}$) are distinct from $\boldsymbol{\theta}$, then $\partial\boldsymbol{\theta}/\partial'\boldsymbol{\beta} = \mathbf{0}$ and the missing-data mechanism is ignorable.

As Little and Rubin (1987) note, if the pattern of missing data is *mar* then, for example, the likelihood that profit data for firm i, m_i^p, is missing does not depend upon the value of profits. Perhaps a more intuitive perspective is to recast the problem in terms of the ability to predict the pattern of missingness. In this respect, for a missing value process that is mar, there is no predictive power in the observed values of m^p.

We now consider an alternate representation of a missing-data mechanism.

Definition 3.2. *oar*

The observed data are observed at random (oar) if for each value of the missing data the conditional probability of the observed pattern of missing data, given the missing data and observed data, is the same for all possible values of the observed data.

If definition 3.2 holds then we can further simplify (3.2) by writing

$$f(\mathbf{Q} \mid \mathbf{M}_\ell, \mathbf{M}_{n-\ell}, \boldsymbol{\beta}) = f(\mathbf{Q} \mid \boldsymbol{\beta}), \tag{3.3}$$

such that the missing-data mechanism is completely independent of both missing and observed data. Note that in (3.3) $\boldsymbol{\beta}$ represents the *unconditional* frequency of missing values, such that $\mathbf{Q} \sim Bin(\boldsymbol{\beta}, n)$. In this instance the missing data are missing completely at random (*mcar*). Extending the example given above, missing profit data in a survey of firms would be *mcar* if the database provider randomly erased profit information for a subset of completed questionnaires.

Before turning to an examination of the principle characteristics of imputation techniques, we comment on the distribution of missing data implied by (3.3). By

utilising this distribution on $f(Q/\beta)$ we noted that this implies that the missing data mechanism is *mcar*, such that for the ith column in Q the distribution is binomial with parameter β_i. That is the missing data mechanism is i.i.d. for any element of Q and that each there is no between variable correlation. Given the likelihood that we might identify *good* firms which have responded to most questions and similarly *bad* firms with non-response across questions, then this assumption of independence across the columns of Q may not be valid. Although in this study we focus on *univariate* missing data processes, in future extensions we plan to extend our analysis to a multivariate framework.

4. Principles of Imputation

In this paper our principal objective is to empirically evaluate the performance of a number of imputation techniques. In this section we outline the basic principles of imputation that will inform the methodologies used in our analysis.

A common feature of a number of commonly used techniques is the posterior predictive distribution of missing values given observed data (*ppdmv).* Using the notation of Section 3 we write this as $f(\mathbf{M}_{n-l}|\mathbf{M}_l)$. As discussed in Section 3 a critical issue is how we construct the predictive distribution and whether the missing data mechanism is ignorable. Depending upon how imputation is carried out we may utilise $f(.)$ to motivate either a Bayesian or classical approach. Rubin and Schenker (1986) emphasise that a Bayesian perspective provides the most natural theoretical framework with which to consider imputation, largely because both parameters *and* data are treated as random quantities. In this context we noted that the imputation of missing data adds an additional component of uncertainty in the construction of the posterior distribution of unknown parameters, denoted $g(\boldsymbol{\theta}|\mathbf{M}_l)$. In the presence of missing data $g(\boldsymbol{\theta}|\mathbf{M}_l)$ is constructed by averaging over the posterior distribution of the missing data given the observed, such that $g(\boldsymbol{\theta}|\mathbf{M}_l)$ is given by

$$g(\boldsymbol{\theta}|\mathbf{M}_l) = \int h(\boldsymbol{\theta}|\mathbf{M}_l, \mathbf{M}_{n-l}) f(\mathbf{M}_{n-l}|\mathbf{M}_l) d\mathbf{M}_{n-l}, \tag{4.1}$$

where $h(.)$ denotes the conditional density of $\boldsymbol{\theta}$ given the *complete* data. This particular approach to handling missing data namely sampling from $f(.)$ to gen-

erate multiple values of $\mathbf{M}_{n-l}$ is now formally referred to as multiple imputation, with the first published work in Rubin (1977). In (4.1) the posterior predictive distribution of missing given observed data plays a central role in the calculation of the posterior distribution for the parameters of interest $\boldsymbol{\theta}$. However, in this study the principal focus is upon the use of observed data to provide estimates for missing data, as opposed to estimates of $\boldsymbol{\theta}$. In this respect the *ppdmv* will play a pivotal role in techniques outlined in Section 5.

The distinction between a Bayesian and classical likelihood-based approach to missing data becomes evident if we compare the data augmentation approach with the EM algorithm. Notice that in (4.1) we take multiple draws from $f(\mathbf{M}_{n-l}|\mathbf{M}_l)$. In contrast the use of EM algorithm in incomplete data problems replaces these multiple draws by the *expected value* of the missing data, conditional on the observed data and current values of parameter estimates. As King, Honaker, Joseph, and Scheve (1998) note, the Bayesian approach preserves the whole distribution of the two estimands - the imputed values and the parameters - whereas the EM approach delivers the single, maximum posterior values. From the perspective of a database provider, there is a choice between filling in missing values using single, averaged, imputed estimates or providing the full set of resampled draws from the posterior distribution.

4.1. Explicit versus Implicit Models

Following Roderick, Little, and Schenker (1995) we make the distinction between *explicit* versus *implicit* models of imputations.[6] The hot-deck method which has been used extensively by the Census Bureau for imputing income in the CPS is perhaps the best known implicit procedure. The hot-deck approach to imputation is based upon selecting data from a distribution of potential donor values for each missing value. For each missing value the approach identifies a matching observed value using a set of covariates that are observed for both respondents and non-respondents. In general the larger the set of observed data the higher the likelihood that any donor value will represent an accurate estimate of the missing value.

[6] A method which combine both implicit and explicit models has been proposed by Little (1998)

Welniak and Coder (1980) utilise the hot deck approach to impute missing values for earnings classifying non respondents into one of eight groups based upon a combination of missing values for earnings (recipiency and amount), work experience and longest job. Thereafter respondents supply donor values based upon similar values of a large number of characteristics such as work experience, sex, age race and region of residence. Using just two categorical variables x_2 and x_3, and following Lillard, Smith, and Welch (1986), we outline the basics of hot deck imputation. Let x_{ijl} represent total employment for firm l in a cell with $x_1 = i$ and $x_2 = j$, such that

$$x_{ijl} = \mu_{ijl} + \varepsilon_{ijl},$$

represents a fully interactive ANOVA model with μ_{ijl} denoting expected employment and ε_{ijl} is a random error term. Assuming firm m is the donor firm, the imputed value x_{ijm} is written as

$$x_{ijm} = \overline{x}_{ijl} + r_{ijm}$$

where $\overline{x}_{ijl}$ is an estimate of μ_{ijl} and $r_{ijm} = (x_{ijm} - \overline{x}_{ijm})$ is a residual for a respondent chosen randomly from the $(i, j, k)th$ cell. Although this approach imposes few conditions, it does assume that the missing data is *mar*, that is within each cell the distribution over the phenomenon of interest is the same for firms responding and non-responding. However, the more observed information that is accounted for, the more likely that any non-response bias can be reduced.[7] Also it is important to note the observation by David, Little, Samuhel, and Triest (1986), namely that nonrespondents with extreme values will be difficult to match.

4.2. Non-Ignorable Non-Response

As discussed in Weeks (2001), a critical feature of the imputation procedure is whether we may reasonably assume that the missing data mechanism is ignorable. In this instance, imputation is considerably easier. Hot-deck and regression based imputation rely on this assumption. Greenlees, Reece, and Zieschang (1982) discuss imputation techniques for non-ignorable missing data processes.

[7] Rubin and Schenker (1986)

The problem of making inferences in models subject to *nonignorable* nonresponse has been treated extensively within the econometrics literature (see, for example, Horowitz and Manski (1998)). Regression based (conditional) imputation is not-valid if, for example, the *dependent* variable is constrained to lie within a given interval. For example, in the case of sampling from a (truncated) distribution where y is observed if $y > \alpha$, then $E(y|\mathbf{x}, y > \alpha) \neq \mathbf{x}\boldsymbol{\beta}$ (even if $\boldsymbol{\beta}$ is unbiased) since the stochastic component is obviously correlated with $\mathbf{x}$ and will not have zero mean. Subsequently imputation based upon the index $\mathbf{x}\hat{\boldsymbol{\beta}}$ will produce biased estimates of the missing values (see Greenlees, Reece, and Zieschang (1982) for further details).

We examine this problem further by considering the following model of missing profits data. Letting $q_i = \mathbf{1}(m_i^p \text{ is missing})$, where $\mathbf{1}(.)$ denotes the indicator function, and as before m_i^p denotes profits for the ith firm, a logistic model of the missing data process may be written as

$$P(q_i = 1|\mathbf{m}_{i,-1}, m_i^p) = [1 + \exp(-\alpha - \theta m_i^p - \boldsymbol{\omega}\mathbf{m}_{i,-1})]^{-1}, \tag{4.2}$$

where $\mathbf{m}_{i,-1}$ denotes a $k \times 1$ vector of covariates, α and θ are unknown scalar parameters and $\boldsymbol{\omega}$ is a $1 \times k$ vector of parameters. The principal problem with (4.2) is that in most cases θ is not identified given that when $q_i = 1$ profits data are missing. However, in this study it is possible to conduct a test of whether or not the missing data process is *mar* by using data for both respondents and non-respondents provided in the ICC series. If either the missing profits or employment data are missing at random then the *mdgp* simplifies to (3.2). If we reject the null hypotheses that $\theta = 0$ then the missing data process is not *mar,* such that imputation based solely on observed data in $\mathbf{m}_{i,-1}$ will generate biased and inconsistent estimates of the missing data. If we cannot reject the joint null that both $\theta = 0$ *and* $\boldsymbol{\omega} = \mathbf{0}$ then the process is *mcar*. Given the availability of a secondary data source providing data values for the non-respondents, we are also able to determine the consequences of these biases for predicting the missing data.

As noted above in most cases it will not be possible to identify θ, and therefore a test of non-ignorable nonresponse is not be possible. However, following the seminal work of Heckman (1976) it is possible to circumvent this problem by

imposing some additional structure on the problem. First, we introduce an unobserved random variable m_*^p which determines whether or not m^p is observed; m_0^p denotes profits for *respondent* firms. The generic form of this problem is known as stochastic censoring and involves specifying the joint distribution of two variables m_0^p and m_*^p,

$$\begin{pmatrix} m_0^p \\ m_*^p \end{pmatrix} \sim BVN\left(\begin{bmatrix} \mathbf{x}_0\boldsymbol{\beta}_1 \\ \mathbf{x}_*\boldsymbol{\beta}_2 \end{bmatrix}, \begin{bmatrix} \sigma_1^2 & \rho\sigma_1 \\ \rho\sigma_1 & 1 \end{bmatrix}\right), \tag{4.3}$$

where $BVN(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ denotes the bivariate normal distribution with mean $\boldsymbol{\mu} = (\mathbf{x}_0\boldsymbol{\beta}_1, \mathbf{x}_*\boldsymbol{\beta}_2)'$ and covariance $\boldsymbol{\Sigma}$. $\mathbf{x}_0$ and $\mathbf{x}_*$ denote, respectively, possibly overlapping sets of covariates, and $\boldsymbol{\beta}_1$ and $\boldsymbol{\beta}_2$ are unknown parameter vectors. Specifically we may think of m_*^p as the net propensity to report profit data such that q may be written as $q = \mathbf{1}(m_*^p > 0)$ where $\mathbf{1}(.)$ denotes the indicator function. Based upon (4.3) and the form of the missing data process, we may write the probability that $m_*^p < 0$, $\Pr(m_*^p < 0|m^p, \mathbf{x}_*)$, as

$$\Phi(\mathbf{x}_*\boldsymbol{\beta}_2) + \frac{\rho\sigma_1(m^p - \mathbf{x}_1\boldsymbol{\beta}_1)}{\sqrt{1-\rho^2}}. \tag{4.4}$$

For $\rho \neq 0$, the missing data mechanism depends on m^p which is only partially observed.

Using (4.3) the conditional population regression function for firms reporting profits is

$$E(m_0^p|m_*^p > 0) = \mathbf{x}_0\boldsymbol{\beta}_1 + \zeta \underbrace{\phi(\gamma)/(1-\Phi(\gamma))}_{T_1}, \tag{4.5}$$

where $\gamma = \mathbf{x}_*\boldsymbol{\beta}_2$, ϕ (Φ) denotes the density (distribution function) of the standard normal distribution, T_1 represents an artificial regressor used to correct for the effect of non-random nonresponse and $\zeta = \rho\sigma$. Note that T_1 represents one form of *self-selection* adjustment for a given *user* of the data. However, from the perspective of a database provider, interested in producing a *complete* set of profit data, the predicted profit values for non-respondents may be *imputed* using

$$E(m_0^p|m_*^p < 0) = \mathbf{x}_2\beta_1 - \zeta\phi(\gamma)/\Phi(\gamma). \tag{4.6}$$

It is important to emphasise that the Heckman procedure is highly sensitive to model misspecification particularly with respect to bivariate normality and

the division of the total set of covariates into sets $\mathbf{x}_0$ and $\mathbf{x}_*$. For example, Olsen (1980) has noted that if $\mathbf{x}_1$ and $\mathbf{x}_2$ coincide, then the model is only identified by the nonlinear transformation on $\phi(\boldsymbol{\gamma})/\Phi(\boldsymbol{\gamma})$. In practical applications it is necessary for $\mathbf{x}_0$ and $\mathbf{x}_*$ to be distinct. However, prior knowledge as to the appropriate set of zero restrictions may be lacking. Note that based upon (4.5) a test of whether profit data may be considered *mar* - is a test of $\zeta = \rho\sigma_1 = 0$. If we cannot reject this null and in addition if $\mathbf{x}_0 \neq \mathbf{x}_*$ then the missing-data mechanism is ignorable such that estimation of the parameters $\boldsymbol{\beta}_1$ and σ^2 based upon an application of OLS on the respondent data alone will generate both unbiased and efficient parameter estimates (for the *end user*), and unbiased estimates of missing data (for *database providers*).

As alluded to by Horowitz and Manski (1998), the principal problem with this approach is that (4.6) is only appropriate if the (generally) untestable assumptions are valid. Subsequently, a number of analysts have evaluated the performance of a number of approximations using a quasi-experimental framework. In an analysis of imputation techniques applied to missing wage and salary data in the Current Population Survey (CPS), Greenlees, Reece, and Zieschang (1982) utilise a secondary data source from the Internal Revenue Service (IRS) to compare imputed data with nonrespondents IRS wage data. The authors find that an approach which utilises a stochastic censoring model represents an improvement over a standard regression approach which assumes that the missing data process is *mar.* Given access to a complete IRS wage series the authors were able to test hypotheses which in most circumstances are not verifiable. For example, a negative and highly significant coefficient on the wage variable in a probability of response mode (i.e. θ in (4.2)) resulted in the rejection of non-ignorable nonresponse. In addition, despite finding approximate symmetry, a large kurtosis value on the residuals from the wage equation resulted in the rejection of the normality hypothesis.

5. Resampling from the Posterior Predictive Distribution of Missing Values

In Section 4.1 we outlined the basics of the hot deck approach to imputation. We also noted that this is an *implicit* method since the manner in which the observed data is partitioned into non-overlapping donor cells does not depend upon a parametric model. An alternative way of organising the observed data is to specify an *explicit*, parametric model of the missing data. Weeks (2001) outlined this approach and we borrow heavily from the exposition, using the same notation as developed in Section 4.2.

We let $\Phi(\mathbf{x}_{*i}\widehat{\beta})$ denote an estimate of the probability that data (in this application either employment or profit) is missing for firm i. Using the quantiles of $\Phi(\mathbf{x}_{*}\widehat{\beta})$ we partition predicted probabilities for the n firms into J equal parts. Within each jth quantile let Mq_j (Oq_j) denote the number of missing (observed) data, such that the total number of observed (missing) values is therefore

$$\sum_{j=1}^{J} Oq_j = n_1; \sum_{j=1}^{J} Mq_j = n_O \quad (n = n_1 + n_O). \tag{5.1}$$

We resample from the posterior predictive distribution of the missing data *conditional* upon the observed data and an explicit model of the missing data process. We do this by drawing, for each jth quantile, a random sample with replacement of size Mq_j from Oq_j, indexing these draws by m and repeating M times. Here we note an important difference between a Bayesian and classical approach to imputation. Within a classical framework an estimate of the missing observation i in the jth quantile is then given by

$$\hat{y}_{ij} = \frac{1}{M} \sum_{m=1}^{M} y_{Oj}^{m}, \tag{5.2}$$

where y_{Oj}^{m} denotes an observed value in the jth quantile. However, the multiple imputations (y_{Oj}) represent independent draws from the posterior predictive distribution - namely $f(y_{n-l}|y_l)$. Therefore, if we are primarily interested in obtaining *estimates* of the missing quantity for non-respondent firms, then reducing the information in a posterior predictive distribution to an average for each quantile, as in (5.2) is, in part, valid. However, independent of any differences in methodology

due to opposing classical and Bayesian perspectives, it is obvious that as providers of data it is important to indicate the degree of uncertainty attached to individual imputations, and thereby differentiate between imputed and fully observed data.

An advantage of the procedure outlined above is that it does not impose an assumption of bivariate normality, nor does it impose a parametric model for the mean equation when imputing missing observations i.e. it does not assume that imputed values are generated by (4.5). Instead differences between respondents and non-respondent firms are controlled for by the specification of a missing data probability model, with the set of potential 'donor' firms stratified based upon the quantiles of $\Phi(\mathbf{x}_*\widehat{\boldsymbol{\beta}})$. In this respect, it is important to note that the use of a resampling approach is predicated on both the specification of a probability model with high predictive power, and the number of quantiles. We return to these operational issues in Section 9.

To examine this procedure more closely we examine the following simplification. First, we consider four types of firms all of which have observed profit data (P), denoting these as Group A $= \{a_i, b_i, c_i, d_i\}$. Based on observed profits for Group A and $\Phi(\mathbf{x}_*\widehat{\boldsymbol{\beta}})$ we write the joint distribution of the two binary classifiers, high/low profits and high/low $\Phi(\mathbf{x}_*\widehat{\boldsymbol{\beta}})$ in Table 1. In the case of firms for which profit data is unobserved we define two firm types which we denote Group B $= \{e_i, f_i\}$. The resampling approach is based upon sampling data from a_i, c_i (b_i, d_i) firms to impute values for e_i (f_i). For example, an estimate of the profit rate for a firm of type e_i may be written as

$$\widehat{P}_{(e_i)} = \frac{1}{M}\sum_{m=1}^{M} P_{(a_i,c_i)}. \tag{5.3}$$

Firms of type e_i (f_i) are matched with donor firms a_i and c_i (b_i, d_i) based upon the probability estimates $\Phi(\mathbf{x}_*\widehat{\beta})$. In the case of e_i firms, firms of type a_i and c_i will contribute profit values to the summation in (5.3).

The principal drawback of this methodology are analogous with the problems encountered with hotdeck imputation, namely that respondents and nonrespondents are assumed to have the same profit distribution within any quantile defined by the missing-data probability model.[8] To see this we note that there exists a

[8]This implies that the 'nonresponse' mechanism is ignorable. As noted above, the better

bound on the predicted values, $\widehat{P}(e_i)$ of the form

$$P^{\min}(a_i, c_i) < \widehat{P}(e_i) < P^{\max}(a_i, c_i) \tag{5.4}$$

imposed by the assumption that the missing data process is *mar*. For example, if the observational rule generating observed profits, P_0, were $P_0 = \mathbf{1}(P < \zeta)P$, such that profits are only reported if they are less than a threshold quantity ζ, then resampling from a lower truncated distribution of observed profits will obviously introduce attenuation bias into the estimates of missing profits.

6. Patterns of Missing Data

The extent to which either an explicit or implicit approach to imputation is appropriate will depend inter alia on the *pattern* of missing data. Figures 1, 2 and 3, adapted from Little (1998), indicate a number of possible patterns. δ denotes a sample-indicator variable with $\delta_i = 1$ if firm i is sampled, 0 if not. z denotes qualitative data on firm characteristics such as industry and location identifiers, which in general, are not subject to item non-response. $\mathbf{X} = (\mathbf{x}_1, ..., \mathbf{x}_k)'$ denotes a $n \times k$ matrix of data, with each $\mathbf{x}_i$, $i = 1, ..., k$, representing a $n \times 1$ vector. For ease of exposition we assume that the primary interest is in obtaining a complete set of observations for $\mathbf{x}_1$, denoting the remaining variables by $\mathbf{X}_{-1}$. Region A denotes information on non-sampled firms.

Figure 1 presents the case where the variable of interest, $\mathbf{x}_1$, has missing data for cases $j, ..., n$, denoted by region B. For $\mathbf{x}_2, ..., \mathbf{x}_k$ there is no item non-response. In this instance these variables may be used as input into either explicit or implicit models of imputation, without the need to worry about missing data in $\mathbf{X}_{-1}$. Figure 2 presents a very different scenario. For cases $j, ..., n$ data is missing for all k variables, again denoted by region B. In this instance neither implicit or explicit techniques represent viable approaches to imputation. Whether or not model-based inference could proceed using the set of complete observations will depend upon the form of the observational rule. For example, we could not ignore

the probability model, the better are we able to control for possible differences between the two groups.

this particular manifestation of questionnaire non-response[9] if, in the case of a business survey, non-respondents were located in a specific region, and that firm characteristics were, in part, geographically determined. The situation between these two cases is represented by Figure 3. This pattern of non response has been referred to as *monotone* or *nested* (see Hartley and Hocking (1971) and Rubin (74)). Here $\mathbf{x}_1$ is observed for j observations, $\mathbf{x}_2$ for $j' > j$, with variables $\mathbf{x}_3$ to $\mathbf{x}_k$ fully observed. In this situation basing imputation on observed data in either $\mathbf{x}_2$ and/or $\mathbf{x}_3$ will create additional problems. For example, consider the case where $\mathbf{x}_2$ is a trinomial categorical variable and is to be used for hot-deck imputation. In this instance, the n observations are stratified according to the realisation of $\mathbf{x}_2$, and cases with observed data in $\mathbf{x}_1$ acting as potential donors for non-respondents. In this instance, missing data on $\mathbf{x}_2$ does not create a problem.

7. Data

7.1. The CBR SME Survey data

The CBR has carried out a series of postal surveys of independent firms in the UK Small and Medium Sized Enterprise (SME) sector, employing less than 500 workers. These surveys have taken place in 1991, 1993, 1995, 1997 and 1999 (see Cosh and Hughes (2000)). The current paper uses data from firms, which responded to both the 1991, and 1995 surveys and includes data reported by those firms in the 1993 survey.

The achieved sample in 1991 was split equally between manufacturing and business services. It included 2028 respondents (a unit response rate of 32.9%). The respondents to the original survey were re-surveyed in 1993. This produced 1341 responses (a unit response rate of 66.1%). As an integral part of the re-survey process, the CBR monitored the status of the original 2028 firms after 1991 using up-dated Dun and Bradstreet information, on-line data from Inter-Company Comparisons (ICC), and microfiche data from Companies House.[10] In

[9] Note that this is not strictly complete questionnaire response since these firms have reported information on qualitative firm characteristics such as industry and location.

[10] The ICC database is a standardised financial accounts database based on the returns submitted by companies to Companies House as part of their statutory reporting requirements.

1995 a further survey was conducted of all those respondents in 1991 which on the basis of this monitoring activity could be identified as still alive and independent in that year. Of the 1592 firms so identified 998 responded (a unit response rate of 62.7%).[11] These 998 firms constitute the panel from which data for the current study is drawn.

Table 1 presents a summary of the areas covered by the 1991, 1993, and 1995 surveys. The 1991 survey was extensive with 61 questions covering eight topics, resulting in 316 variables.[12] The focus of the second survey was to update selected performance variables and to examine in greater detail than in 1991 the financing constraints facing UK SMEs; this resulted in 11 questions and 30 variables. The 1995 survey included 29 questions yielding 198 variables. Each survey included comparable questions on the turnover, employment, and profits size of the respondent.

7.2. The Amalgamated CBR-ICC Dataset

In this paper we construct a quasi experimental framework and examine the performance of a number of imputation methods applied to the problem of non-response in business survey data. We adopt a methodology that exploits the availability of a secondary data source in the form of a comparable employment and profit series through data which companies must deposit at Companies House. Subsequently it was necessary to match each CBR survey case in the panel with the corresponding set of company accounts on the ICC database.

The ICC database contains a wide range of standardised financial and accounting ratios as well as additional information relating to credit worthiness, such as the presence, or absence, of County Court Judgements affecting the firm. It was possible to attach ICC accounts to 965 firms in the panel of 998 responding in

[11] See Bullock, Duncan, and Wood (1996).

[12] The 1991 survey used an employment-size stratified sampling design. The sampling framework used in the construction of the survey was the Dun & Bradstreet (D&B) database. This database has its origin in the credit-rating business and as a result is probably biased towards the inclusion of a relatively higher proportion of expanding firms seeking finance than is true of the enterprise population as a whole. It is also known, that at that time it under-represented sole proprietors, partnerships and single person self-employed enterprises compared to the overall enterprise sector.

both 1991 and 1995, the balance being partnerships proprietorships or companies with untraceable ICC accounts. The precise financial data available for each of the 965 companies is, however, affected by the extent to which these companies availed themselves of the right to submit modified or abbreviated accounts to Companies House. For example, the Companies Act 1985, and subsequent amendments, defined classes of small and medium sized companies which may submit *modified* accounts which did not need to include, in particular, data on profits, turnover, or employees. In 1995, a substantial proportion of the CBR database could claim exemption (other years) and submit such modified accounts on the basis of satisfying any two of the following three criteria at the time of the 1995 survey:[13]

	Small	Medium-Sized
Annual turnover not exceeding	£2,800,000	£11,200,000
Balance sheet total not exceeding	£1,400,000	£5,600,000
Average number of employees not exceeding	50	250

These basic size cut-offs are the subject of periodic amendment (for instance the DTI issued a consultative document in May 1995 which proposed to increase the limits further by 50 per cent). Even in 1995, however, a substantial proportion of the CBR database could claim exemption. Thus over 600 of the panel sample employed less than 50 workers and 50% had a turnover of less than £1million.[14]

For the purposes of this study, therefore, two sub-samples of firms were formed. These consisted, respectively, of firms for which there were no item missing values for ICC profits on the combined CBR-ICC database, or for which there were no item missing values for ICC reported data (either profits or employment) on the combined CBR-ICC database. One of each of these samples was formed for the CBR survey data year of 1991 which yields the largest sample sizes for analysis.

[13] Source: The Coopers and Lybrand Manual of Accounting Accountancy Books, Institute of Chartered Accountants of England and Wales, November 1995 pp 36001-36007.

[14] See Cosh, Duncan, and Hughes (1996).

7.3. Patterns of Missing Data: Profiling the Non-Respondents

In order to establish a profile of firms who were non-respondents on either profit or employment data in 1991 we compared the values of a wide range of CBR survey variables for those cases where CBR profit and employment data were missing. This was done for the whole sample and for each of the two sub-samples. Table 3 presents an analysis, for the whole sample, of all companies which had an ICC account attached (whether modified or not), comparing those cases where CBR *employment* data was missing and those where it was present. Table 4 provides a similar analysis but this time comparing cases where CBR *profits* data were missing with firms where they were not. Row 1 of Table 3, for example, shows that of the 198 firms reporting in the 1991 survey on additional finance obtained from HP or Leasing companies, (ADDFIN13) 40 had missing values for CBR employment. The mean percentage of finance obtained from this source by those with missing (non-missing) employment data was 23.6% (17.4%). This difference was statistically significant using the non-parametric Mann-Whitney Test but was not using a standard 't' test.

Tables 5 and 6 repeat the exercise but this time on two sub-samples where ICC employment and profit data respectively were present.[15] Thus row 1 of Table 4 shows that in the sub-sample of cases where ICC employment is always present, CBR employment data was missing for just 7 cases, and present for 111 cases, of the total of 118 cases where the respondent reported that additional finance was obtained from venture capitalists. The difference in the proportion of finance obtained from this source between the missing and non-missing CBR employment data group was significant on the parametric test but not on the non-parametric test.[16]

The CBR variables presented in each table are a subset of a wider range of CBR variables, which on either theoretical or empirical grounds could be used

[15] Note that in Tables 3 and 4 the two sub-samples were selected on the basis that an ICC account was attached. The statistics in Tables 5 and 6 are restricted sub-samples based upon the actual reporting of, respectively, ICC employment and profit data.

[16] In the case of employment data we found that there were very few firms with both matching ICC employment data *and* missing CBR employment. Subsequently at this juncture we decided not to attempt imputation, and simply focus on the profit data.

as predictors of employment or profit. The vast majority of the variables tested relate to 1991. Only those variables where statistically significant differences emerged between the respective missing and non-missing groups (using either the parametric or non-parametric test) are reported.

The sub-set of variables which emerged as significant varies from sample to sample. There is, however, some pattern of differences between the two employment based samples (Tables 3 and 5), and the two profit based sample (Tables 4 and 6). In the latter for instance the profit related variables relating to overdraft and interest payment patterns emerge as discriminators between the missing and non-missing groups. They do not, however, appear in the employment related samples where their discriminatory power is insignificant and they are not reported.

7.4. Matching CBR and ICC Data

Although we have referred to the ICC database as a secondary data source providing observations for firms classified as non-respondents in the CBR data, there are two important qualifications. First, as noted above, there does not exist a match for every non-respondent in the CBR database given that it is not compulsory for all firms to provide account information. Second, for reasons discussed in Section 7.2 this data does not always provide a perfect match. David, Little, Samuhel, and Triest (1986) note similar problems in a study of comparing different methods of income imputation, with differences in the inclusion of sick pay across data sources.

To examine the relationship between the two employment series we estimated a simple piece-wise linear regression of the form

$$CbrE_i = \sum_{j=1} \mathbf{1}(\alpha_j < IccE_i < \tau_j) + \beta_1 IccE_i + \beta_2 IccE_i^2 + \varepsilon_i, \qquad (7.1)$$

where i indexes firms for which matched data exists, $CbrE_i$ and $IccE_i$ denote, respectively, CBR and ICC employment figures, $\alpha_j = \{0, 7, 20, 50\}$ and $\tau_j = \{6, 19, 49, \infty\}$ represent employment thresholds such that $\sum_{j=1} \mathbf{1}(.)$ denotes a series of mutually exclusive dummy variables representing mean adjustments.[17] Param-

[17] Note that these thresholds were selected based upon inspection of the respective distribu-

eter estimates are given in Table 7. A number of observations are noteworthy. First, the mean adjustments provided by the series of employment thresholds, α_j, τ_j indicate that the CBR measure of employment exceeds that of the ICC with the difference increasing with the level of employment. With the adjustment provided by (7.1) the two series are reasonably comparable.

We now turn to the problem of matching CBR and ICC profit data. In Table 8 we present summary statistics for the two profit series. Of the 243 firms with reported ICC profits, 181 firms also report profit data in the CBR database; 62 firms fail to report profit data. Although both the CBR and ICC profit distributions include both negative and positive profits, a simple visual inspection might suggest that the missing data process is not random, in the sense that for the ICC profit data which effectively defines the population of interest, the mean is negative, compared to a positive mean for the CBR series. Thus, we might conclude that the non-respondents in the CBR database are those with very low (and negative) profits. However, this conclusion is only possible if the two profit series are directly comparable. Inspection of summary statistics for *matched* firms in Table 8 demonstrates that the two series are not comparable, with, for example, CBR (ICC) profits exhibiting positive (negative) skewness. Based upon this observation it is therefore difficult to reconcile the two series. In Table 9 we report the results from an attempt to do this by estimating a piece-wise linear regression of the form

$$\begin{aligned} CbrP_i &= \sum_{j=1} \mathbf{1}(\alpha_i < IccP_i < \tau_i) + \theta_1 IccP_i + \theta_2 IccP_i^2 + \\ &\quad \theta_3 IccP_i^3 + \varepsilon_i, \end{aligned} \tag{7.2}$$

where $CbrP_i$ and $IccP_i$ denote, respectively, CBR and ICC profits, and $\alpha_i = \{-\infty, -249, -99, 0, 200, 400, \infty\}$, $\tau_i = \{-250, -100, 0, 0, 1, 201, 401\}$ are mean adjustments. Parameter estimates, presented in Table 9, reveal that there are some significant differences in the way in which profits are recorded in the CBR and ICC databases. As an example, firms with $200 < ICC$ profits[18] $<= 400$ the required mean adjustment for comparability is -245. Of particular note is the

tions of ICC and CBR employment data.

[18] Profits are measured in £000.

threshold coefficient $\alpha_3\tau_3$ which indicates that when zero ICC profits are recorded CBR profits are on average 120. Relative to the employment data we observe a poor fit with only 60 per cent of the variation in the CBR employment data explained by the ICC data.[19] Although this obviously represents a problem for evaluating the relative performance of imputation methods using methods such as mean squared error, this problem does not affect the *ranking* across methods. This follows since imputation error will be composed of a method specific component and a constant reflecting the fact that the adjusted secondary data source, here ICC profits, is not directly comparable with the CBR profit series.

8. Results

Below we present the results from applying a number of imputation methods to the problems of missing profit data in the CBR Small and Medium sized enterprise database. We initially present the results of estimating a number of models which will provide *forecasts* of the missing data. Following this we evaluate the differential performance of our imputation techniques using the ICC profit series as a benchmark.

We begin with an examination of the performance of a simple conditional mean model of imputation which assumes that the missing data generating process is *mar*. In Table 10 we present the results of estimating a simple linear model of profits, using the regressors: turnover (TURN), age of the chief executive (CEAGE1), a binary variable reflecting skill shortage (SKILL), a binary variable indicating whether formal training is provided (TRAIN) and LIMAB101.[20] We denote this set by $\mathbf{x}$. It is important to note that this procedure takes no account of the possibility that the *observed* CBR profit data represents a set of non-random responses, insofar as, for example, that the *mdgp* is determined in part by the level of profits. In addition, the set of parameters in Table 10, say $\boldsymbol{\beta}$, are estimates based upon a set of firms where the intersection of non-missing data over $\mathbf{x}$ and CBR profits is not empty.

[19] See Appendix 1 for a full discussion of the reasons for differences between CBR and ICC data.

[20] See Appendix 3 for a complete set of definitions.

As discussed in Section 4, a model that approximates the missing data generating process is used in two imputation techniques used in this paper: a conditional mean regression which corrects for non-random non-response; and the multiple imputation approach which is based upon resampling from the predictive posterior distribution of the missing data. In addition, since we have access to a secondary data source in the form of the ICC profit data, we may also test for whether such a correction is warranted. In Table 11 we present the results of estimating two models. In columns 1 and 2 we present parameter estimates and t-ratios based upon estimating a simple probit missing data model. We note that by including the ICC measure of profits as a regressor we are able to test whether or not the *mar* assumption is valid. We reject the *mar* null at the 6% level and observe that as profits decrease there is a statistically significant reduction in the likelihood of reporting profit data. We also emphasise the significance of the variable TOTMISS91. This variable is an ordinal integer variable which records, for each firm, the number of variables which have missing observations across a relatively large set of firm characteristics. Although there is the obvious issue that this variable is unlikely to be exogenous, the highly significant effect indicates that both description and inference within the confines of missing data models needs to be conducted in a multivariate framework.

Given that in most instances analysts will not have secondary data series such as the ICC profit series, in columns 3 and 4 of Table 11 we present the results of estimating the same model of missing data but placing a zero restriction on ICCP91. The predicted probabilities ($\widehat{p}$) will be used in the construction of a conditional mean imputation model with a correction for the missing data observational rule, and also in a Multiple Imputation procedure where $\widehat{p}$ will be partitioned using quantile thresholds such that, for each quantile, observed profit records can be used as donors for missing profits.

In Table 12 we present estimates of the same model as in Table 10 but with a correction for non-random non-response for profits.

9. Comparing Imputation Performance

In Tables 13 and 14 we evaluate the performance of the conditional mean (CM), Hot-Deck and Multiple Imputation (MI) techniques. We summarise our results using three criteria: $\Sigma_{IMP}/\Sigma_{OBS}$ is the ratio of the summed imputed to *observed* values, and measures bias at the aggregate level; *MABS* is the mean absolute error between observed and imputed data and computes a disaggregate measure of closeness over firm level differences between imputed values and the observed secondary data; and *Mean* is the simple mean error, again computed over firm level differences. For each criterion *Adj* denotes predicted values using the matching equation (7.2), and therefore represents the adjusted ICC profit data.

Table 13 presents the result for CM and HD. The Hot Deck procedure was conducted on the basis of sorting respondents and non-respondents into imputation classes using one or more variables. Donor values from matching respondents are then used to impute missing values.[21] Given that we are *obviously* unable to sort firms on the basis of a fully observed set of response on profits, HD assumes ignorable non-response. Results I-IV are based upon the use of different sets of firm characteristics. In I we use a single characteristic, TRAIN, a binary variable indicating whether or not formal training is provided. In previous work (see Cosh, Hughes, and Weeks (2000) it has been demonstrated that the provision of training is, alongside other characteristics, a determinant of firm profits and employment growth. However, and as expected, the use of the binary variable TRAIN in isolation provides very little discrimination in terms of the ability to impute missing profit data. In II we perform a hierarchical sorting, using both TRAIN and GROWTH, a nominal variable measuring a firms' planned growth objectives[22]. For both the aggregate ($\Sigma_{IMP}/\Sigma_{OBS}$) and *Mean* measure the imputation performance increases considerably, implying that the additional level of discrimination provided by GROWTH generates a better set of matches. In III we reverse the order, sorting first using GROWTH and then TRAIN. Here we note that order of sorting matters, in that the imputation performance improves considerably rela-

[21] We note that if there is more than one matching respondent for any given non-respondent, we randomly select a value from this subset.

[22] See Appendix 2.

tive to III. A likely explanation for this difference follows from the argument that GROWTH is a better *predictor* of profits than TRAIN. In addition, whereas TRAIN is binary, GROWTH is multinomial and, ceteris paribus, will most likely generate better matches. Note that if the two characteristics produce the same percentage of explained variance, then the order of sorting within a HD imputation would have no effect. In Results IV we experiment with a 3 variable HD procedure using EXPORTS, FINANCE and TRAIN, noting that the results are, with the exception of I, the most disappointing.

Under the heading *Regression* we present the results from two applications of conditional mean imputations. First, we assume that the missing data generating process *(mdgp)* is *mar,* following this by a correction for non-random non response. We also note that given we have available the secondary data source in the form of ICC profits, we were able to reject the null that the *mdgp* was *mar*. Although it is worth emphasising that relative to the *mar* conditional mean model, we observe a substantial fall in *MABS* and a much better correspondence between the ratio of total imputed to observed profits, the absolute level of *MABS* is still high.

In Table 14 we present results for 2 variants of the Multiple Imputation (MI) procedure: first, using deciles to partition the predicted probabilities, and second using quinquartiles, which uses threshold values such that there are 20 bins from which to resample. The reason for choosing these two designs follows from the maintained assumption of the MI method, namely that for the variable in question (in this instance profits), the distribution for respondents and non-respondents is the same within any given quantile. Therefore, and as an example, it may be the case that for a particular decile there exists considerable differences between the two distributions. In this instance by averaging over observed resampled profit data to impute missing values we are obviously using an inappropriate distribution. In contrast, although the variant based upon quinquartile users a finer grid, it is likely that in certain instances the variance around the imputed value may be low, if there are only a few observed records in a particular quantile.

In relative terms the results clearly demonstrate that Multiple Imputation using the quinquartiles is superior over the decile variant. However, in absolute terms the procedure is obviously problematic dependent upon the loss function of

the analyst. For example, positive and negative imputation errors are negligible using 1000 draws from the predictive posterior distribution of missing data. Likewise the ratio of $\Sigma_{Imp}|\Sigma_{OBS}$ converges to one. However, there are still significant imputation errors if we use the *MABS* measure of performance.

10. Conclusion

This paper represents a *preliminary* investigation into the performance of a number of imputation techniques applied to missing profits data in the CBR Small and Medium Sized Enterprise database. Although, the use of a secondary data source providing an observed profit series for a subset of firms with missing CBR profit data, provides a necessary benchmark, one obstacle to date has been the problem of reconciling the two series. Subsequently in evaluating performance across techniques, it is perhaps better at this juncture to focus upon the *ranking* of different techniques, given that, for example, the absolute level of mean absolute imputation error contains both real error and a term (constant across the different techniques) representing the lack of comparability between the two series.

In this regard the Multiple Imputation (MI) procedure represents the most encouraging results especially when a large number of resamples are used. By comparing the MI procedure using deciles as opposed to quinquartiles to partition the posterior distribution of missing values, we also highlighted the importance of a number of operational aspects which are often glossed over in the literature. In future work we plan to examine these issues in more detail.

We emphasise that this paper has provided a framework for ongoing research. In this respect we note the following key points. First, although we believe that the rankings of imputation methods revealed in this research are, in general, likely to be preserved in subsequent work, we hope to be able to reduce the level of mean absolute error by: a) a better reconciliation of the observed and secondary data sources; and b) better models of the phenomenon under investigation. In this report we view the current set of results as demonstrative rather than definitive. Although our interest is in minimising imputation error and thereby constructing a complete dataset with reliable estimates of missing values, we obviously need to

spend more time in searching for better models of profits and probability models for missing data. Second, the missing data analysis and imputation has been conducted within a *univariate* framework. This is obviously a disadvantage insofar as univariate missing data processes are correlated[23]. In addition although it is easy to justify our focus upon a single firm attribute such as profits, we have encountered problems in our analysis given the incidence of missing data in variables used to either *match* or *predict* profits for nonrespondent firms. In future research we intend to explore multivariate models of missing data.

[23]This was partially revealed in Section 8 when we used the ordinal variable TOTMISS91 as a covariate in the missing data model for profits.

Appendix 1

Reasons for differences between CBR and ICC data on profits and employment

One simple reason for differences in both the profit and employment data between ICC and CBR is that the reported data relate to different accounting year-ends. It is also possible that ICC employment data relate to averages over financial years whilst CBR employment data relate to estimates at a point in time.

In the 1991 CBR survey the data on profits and employment were requested without requiring the companies to state the accounting year-end to which they related. In this study we compare ICC data for accounts ending in the financial year April 1990 to March 1991 with data from the 1991 CBR survey. This should correspond to the period the survey firms report on since the CBR survey took place between April and September 1991. It is possible, however, that the CBR data relates to earlier accounting years. This could bias the CBR employment data downwards relative to the ICC data in so far as the average company grew over the period 1990-1991.On the other hand, in these circumstances, the CBR employment data could equally be biassed upward. This could happen if respondents reported employment at the time of the survey, instead of reporting their average employment over the last financial year, (as required in the note to the profit and loss accounts of companies submitting full returns to companies House). Profits could be downward or upward biassed by differences in accounting year ends depending upon patterns of profit movement over time in the affected firms. Any upward bias would be compounded given inflation over the period analysed.

There are a number of other reasons why the data in the CBR survey may differ from those reported in the company accounts database. One possibility is clerical or transcribing error. There is a high premium attached to accuracy in the preparation of a commercial database such as ICC, which is used, for credit checks and credit scoring. This means the likelihood of error should be small but we have no external checks on this. The CBR data is subject to a number of rigorous internal consistency and screening tests. It has, moreover, been used extensively in econometric analysis of performance and other SME characteristics where substantial data checking has also been carried out. Errors may nonetheless

remain but we have no reason for expecting any systematic differences between the CBR and ICC datasets.

Another possibility affecting profits is that there may be an upward bias in the CBR dataset. Profits may be inflated for reasons of reputation in the CBR returns and losses may be concealed in a way that should not be possible in audited returns to Companies House. Further, the CBR profits question requests data on pre-tax profits prior to the deduction of interest payments and *prior to the deduction of directors' emoluments.* It is not possible to derive this measure exactly from the ICC database since Directors' emoluments are not available in a form, which can be added back to pre-tax, pre-interest profit. The ICC series should therefore be downward biassed.

Appendix 2

All of the computations were carried out using the Ox programming language[24]. With the exception of the multiple imputation procedure outlined in Section 5, the computations involved were straightforward. Below we present the code for the *MultiIMPUTE* function.

```
MultiIMPUTE (OrVarIMP, PostProbMiss, PrbQuant, Miss, nIMPUT
```

Objective:

Calculate Posterior Distribution of Missing Data and use the calculated quantiles to partition the predicted probabilities facilitating the (multiple) imputation of missing data.

Arguments:

`OrVarIMP` = Original Variable to impute missing values
`PostProbMiss` = Posterior Predictive Density of Missing Data
`PrbQuant` = Quantiles
`Miss` = 0,1 variable; 1 if Missing
`nIMPUT` = number of imputations

Notation

`q` = Observed and Missing Data in kth Quantile (k x 1)
`qO` = Observed CBR Data in kth Quantile (k x 1)
`qM` = Missing CBR Data in kth Quantile (k x 1)
`cntOq` = number of obs in kth Quantile (k x 1)
`cntMq` = number of miss in kth Quantile (k x 1)
`qrindex` = observation number of Missing value

Return : IMPVar

[24] see Doornik (1999).

```
MultIMPUTE(OrVarIMP,PostPROBMiss,PrbQuant,Miss,nIMPUT)
{
decl i,j,k,cntMqx,cntMq,cntOq,cntOqz,q,qO,qOx;
decl obsNUM,ranINT,qrindex,MultIMP,MultIMPx,means,stdevs;
decl IMPVar = OrVarIMP;
obsNUM = range(1,rows(OrVarIMP),1)';

// OUTER LOOP: individual quantiles
for(k=0,MultIMPx=<>,cntMqx=<>,cntOqz=<>,qrindex=<>,means=<>,
stdevs=<>;k<rows(PrbQuant)-1;++k)
{
q = selectifr(OrVarIMP,PostPROBMiss .> PrbQuant[k] .&& PostPROBMiss
.<= PrbQuant[k+1] );
qO = selectifr(OrVarIMP,PostPROBMiss .> PrbQuant[k] .&& PostPROBMiss
.<= PrbQuant[k+1] .&& Miss .== 0);
qrindex |= selectifr(obsNUM,PostPROBMiss .> PrbQuant[k] .&& PostPROBMiss
.<= PrbQuint[k+1] .&& Miss .== 1);

cntMq = rows(q) - rows(qO);
MultIMP = zeros(nIMPUT*cntMq,1);
cntOq = rows(qO);
qOx = qO~(range(1,rows(qO),1))';
// For each quantile initialise a matrix with M rows and cntMq columns
// M rows for M imputations; cntMq = number of missing per quantile
for(i=0;i<rows(qO);++i)
for(j=0;j<nIMPUT*cntMq;++j)
{
ranseed(123*i*j*k);
ranINT = ceil(ranu(1,10).*rows(qO)); //sampling with replacement
if (qOx[i][1] - ranINT[0] == 0 && cntMq > 0)
 {
MultIMP[j] = qOx[i][0];
```

```
}
}
if (cntMq > 0)
{
MultIMP = shape(MultIMP,nIMPUT,cntMq);
means |= meanc(MultIMP)';
stdevs |=sqrt(varc(MultIMP))';
MultIMPx~= MultIMP;
}
cntMqx~=cntMq; cntOqz~=cntOq;
} // END of MAIN LOOP//
print("\n", " Multiple Imputations Summary Statistics \n");
print("%c",{"Obs # "," Mean ", "Std.  Deviat "},
qrindex~means~stdevs,"\n");
for(i=0;i<rows(OrVarIMP);++i) for(j=0;j<rows(means);++j)
{
if (obsNUM[i] - qrindex[j] == 0) IMPVar[i] = means[j];
}
return IMPVar;
}
```

Appendix 3

CAGE:	chief executive age in years (1991 survey)
COMPS:	number of serious competitors
TURN:	turnover (£000)
EXP:	exports (£000)
LARGEST:	% of sales to largest customer (1991 survey)
PROF:	pre-tax profits (losses) before deduction of interest, tax, and Directors', Partners' or Proprietors' emoluments.
FINANCE:	attempts to obtain additional finance (i.e. in addition to internal cash flow). Yes/No
LIMAB101:	there is a finance limitation in plans for expansion. Yes/No
SKILL:	binary variable equals 1 if the firm is finding it difficult to recruit in certain skill categories, and zero other (1991 Survey)
TRAIN:	binary variable equal 1 if firm provides formal training
GROWTH:	growth objectives over the next 3 years 1 = become smaller; 2 = stay same size; 3 = grow moderately 4 = grow substantially.

ADDFIN12	Additional finance from: Venture capital firms (1 = Yes; 0 = No)
ADDFIN13	Additional finance from: Hire purchase or leasing firms (1 = Yes; 0 = No)
ADDFIN14	Additional finance from: Factoring/invoice discounting firms (1 = Yes; 0 = No)
ADDFIN15	Additional finance from: Trade customers/suppliers (1 = Yes; 0 = No)
ADDFIN17	Additional finance from: Other private individuals (1 = Yes; 0 = No)
ADVICE11	Advice from the Small Firms Service (1 = Yes; 0 = No)
ADVICE13	Advice from DTI's enterprise initiative (1 = Yes; 0 = No)
AVEMP1	Employment in 1990
CE11	Years with Company of Chief Executive
CE14	Is Chief Executive the founder of the Company? 1 = Yes; 0 = No
CEAGE1	Age of Chief Executive
COAD101	Competitive advantage: Price. Measurement scale 0-9 (0 - completely insignificant and 9 = highly significant)
COAD102	Competitive advantage: Marketing and promotion skills. (Measurement scale as COAD101).
COAD103	Competitive advantage: speed of service. (Measurement scale as COAD101)
COAD104	Competitive advantage: Established reputation. (Measurement scale as COAD101)
COAD107	Competitive advantage: Product quality. (Measurement scale as COAD101)
CUS12	Customer collaboration to: expand expertise/products
CUS14	Customer collaboration to: improve finance and market credibility
CUS16	Customer collaboration to: gain access to new equipment/information.
EXP1	Exports (£000s) in 1990
FINANCE1	Attempted to obtain external finance (1 = Yes; 0 = No)
GROWTH1	Your growth objectives: (1 = grow smaller; 2 = stay same size; 3 = grow moderately; 4 = grow substantially)

INNOV15	Introduce any innovations in administration and office systems (1 = Yes; 2 = No)
LIM111	Source of limitation on meeting business objectives: Increasing competition (1 = Yes; 0 = No)
LIM104	" " : Management skills (1 = Yes; 0 = No)
LIM105	" " : Marketing and sales skills (1 = Yes; 0 = No)
LIM106	" " : Acquisition of technology (1 = Yes; 0 = No)
MISEMP1	Binary Variable: 1 if employment data
MISPROF1	Binary variable: 1 if profit data is missing.
OVRDFT11	Do you have over facilities? (1 = Yes; 0 = No)
PARTARR1	Entered into partnership arrangements? (1 = Yes; 2 = No)
PROC1A	Have you recently introduced any process innovation? (1 = Yes; 0 = No)
PARTN11	Partnership arrangements with suppliers? (1 = Yes; 0 = No)
PARTN12	Partnership arrangements with customers? (1 = Yes; 0 = No)
PROD1A	Major innovations in products and services? (1 = Yes; 0 = No)
RD14	No. of full-time staff in Research and Development
SALES11	Percent of sales direct to retail/wholesalers
SALES13	Percent of sales direct to other firms
SALES14	Percent of sales direct to local/central government
SYSSTRD13	Systematic Research and Development into new services (1 = Yes; 0 = No)
TOP15	% of sales to largest 5 customers
TOTFP1	Total number of employees
TURN1	Turnover £(000s) in 1990

References

Ahn, H., and J. Powell (1993): "Semiparametric Estimation of Censored Selection Models with a Nonparametric Selection Mechanism," *Journal of Econometrics*, (58).

Bullock, A., J. Duncan, and E. Wood (1996): "The Survey Method, Sample Attrition and the SME Panel Database," in *The Changing State of British Enterprise: Growth Innovation and Competitive Advantage in Small and Medium Sized Firms 1986-95*, ed. by A. D. Cosh, and A. Hughes, ESRC Centre for Business Research. University of Cambridge.

Cosh, A., A. Hughes, and M. Weeks (2000): "The Impact of Training on Business Employment Growth.," Research Report RR 245. Department for Education and Employment, Sheffield, December.

Cosh, A. D., J. Duncan, and A. Hughes (1996): "Size, Age Survival and Employment Growth," in *The Changing State of British Enterprise: Growth Innovation and Competitive Advantage in Small and Medium Sized Firms*, ed. by A. D. Cosh, and A. Hughes, ESRC Centre for Business Research. University of Cambridge.

Cosh, A. D., and A. E. Hughes (2000): "British Enterprise in Transition: Growth Innovation and Public Policy in the Small and Medium Sized Enterprise Sector 1994-1999," ESRC Centre for Business Research, University of Cambridge.

David, M., R. J. Little, M. Samuhel, and R. Triest (1986): "Alternative Methods for CPS Income Imputation," *Journal of the American Statistical Association*, 81(393), 29–41.

Doornik, J. A. (1999): *Ox: An Object Orientated Matrix Programming Language*. Timberlake Consultants Ltd, London.

Golan, A., G. Judge, and D. Miller (1996): *Maximum Entropy Econometrics: Robust Estimation with Limited Data*. John Wiley and Sons.

Greenlees, J., W. Reece, and K. Zieschang (1982): "Imputation of Missing Values When the Probability of Response Depends on the Variable Being Imputed," *Journal of the American Statistical Association*, 77(378).

Heckman, J. (1976): "The Common Structure of Statistical Models of Truncation, Sample Selection and Limited Dependent Variables, and a Simple Estimator for Such Models," *Ann. Econ. Soc. Meas.*, 5, 473–492.

——— (1979): "Sample Selection Bias as a Specification Error," *Econometrica*, 47, 153–161.

Heckman, J., and V. Hotz (1989): "Choosing Among Alternative Nonexperimental Methods for Estimating the Impact of Social Programs: The Case of Manpower Training," *Journal of the American Statistical Association*, 84, 862–874.

Heckman, J., and J. Smith (1995): "Assessing the Case for Social Experiments," *Journal of Economic Perspectives*, 9(2), 85–110.

Horowitz, J., and C. Manski (1998): "Censoring of outcomes and regressors due to survey nonresponse: Identification and estimation using weights and imputations," *Journal of Econometrics*, 84, 37–58.

Hurd, M., D. McFadden, H. Chand, L. Gan, A. Merrill, and M. Roberts (1997): "Consumption and Savings Balances of the Elderly: Experimental Evidence on Survey Response Bias," Working Paper, Department of Economics, University of California, Berkely.

King, G., J. Honaker, A. Joseph, and K. Scheve (1998): "Listwise Deletion is Evil: What to Do About Missing Data in Political Science," Department of Government, Harvard University.

LaLonde, R. (1986): "Evaluating the Econometric Evaluations of Training Programs with Experimental Data," *American Economic Review*, 76, 604–620.

LILLARD, L., J. P. SMITH, AND F. WELCH (1986): "What Do We Really Know About Wages? The Importance of Nonreporting and Census Imputation," *Journal of Political Economy*, 94(3), 489–506.

LITTLE, R. (1982): "Models for Nonresponse in Sample Surveys," *Journal of the American Statistical Association*, (78), 378.

LITTLE, R., AND D. RUBIN (1987): *Statistical Analysis with Missing Data*. John Wiley, New York.

LITTLE, R. J. A. (1998): "Missing Data Adjustments in Large Surveys," *Journal of Business and Economic Statistics*, 6, 287–301.

RODERICK, J., A. LITTLE, AND N. SCHENKER (1995): "Missing Data," in *Handbook of Statistical Modeling for the Social and Behavioural Sciences*, ed. by G. Arminger, C. Clogg, and M. E. Sobel, pp. 39–75. Plenum Press, New York.

RUBIN, D., AND N. SCHENKER (1986): "Multiple Imputation for Interval Estimation From Simple Random Samples With Ignorable Nonresponse," *Journal of the American Statistical Association*, 81(394), 366–.

RUBIN, D. B. (1977): "Formalizing Subjective Notions About the Effect of Nonrespondents in Sample Surveys.," *Journal of the American Statistical Association*, 72(359), 538–543.

WEEKS, M. (2001): "Methods of Imputation for Missing Data," Mimeo, Faculty of Economics and Politics, University of Cambridge.

WELNIAK, E. J., AND J. F. CODER (1980): "A Measure of the Bias in the March CPS Earnings Imputation Scheme," *Proceedings of the Survey Research Methods Section, American Statistical Association*, pp. 421–425.

Table 1. Joint Probability of Low/High Profits and Low/High Probability of Missing Data.

		$\Phi(\hat{\gamma})$ Low	High
P_i	Low	a_i	b_i
	High	c_i	d_i
		e_i	f_i

Table 2. Coverage of the 1991 and 1993 CBR surveys

	No. questions (No. variables)					
	1991		1993		1995	
Topics						
General business characteristics	6	(17)	5	(9)	5	(11)
Workforce and training	5	(55)	1	(2)	1	(15)
Innovative activity	4	(38)	0	(0)	11	(89)
R&D and other innovation expenditure	2	(5)	0	(0)	3	(18)
Commercial activity and competitive situation	20	(117)	0	(0)	7	(45)
Finance	6	(19)	4	(18)	2	(20)
Executive structure	12	(43)	0	(0)	0	(0)
Acquisition activity	6	(22)	1	(1)	0	(0)
Total	61	(316)	11	(30)	29	(198)

Table 3. A comparison of CBR survey variables for companies grouped by missing/non missing CBR employment data.

	Numbers		Mean		Significance Level	
CBR Variable	Missing	Non Missing	Missing	Non Missing	Mann Whitney 'u'	Students 't'*
ADDFIN13	40	518	23.63	17.35	0.096	0.235
ADDFIN17	40	518	0.00	1.27	0.132	0.000
CE14	60	849	0.58	0.76	0.003	0.011
COAD101	59	854	6.51	5.48	0.019	0.017
COAD103	61	875	7.28	6.52	0.228	0.016
COAD104	61	871	7.46	6.86	0.318	0.030
CUS16	14	267	0.21	0.17	0.078	0.317
FINANCE1	57	878	0.74	0.62	0.088	0.070
LIM106	52	764	2.83	2.17	0.054	0.230
MISPROF1	63	902	0.57	0.80	0.000	0.001
PROC1A	39	539	0.28	0.45	0.047	0.037
PROD1A	44	638	0.41	0.58	0.024	0.029
SALES11	58	874	26.31	31.72	0.047	0.237
SALES14	58	874	9.53	7.09	0.011	0.355
TOP15	54	842	3.20	3.50	0.069	0.081
RD14	34	532	0.26	0.60	0.989	0.037
TURN1	22	851	772.55	1179.83	0.468	0.010

* Equal variances not assumed.

Table 4. A comparison of Values of CBR Survey Variables for Companies grouped by missing/non missing CBR profits data.

	Numbers		Mean		Significance Level	
CBR Variable	Missing	Non Missing	Missing	Non Missing	Mann Whitney 'u'	Students 't'*
ADDFIN14	105	453	1.88	4.29	0.140	0.079
ADDFIN17	105	453	0.24	1.40	0.244	0.001
ADVICE13	208	757	0.21	0.37	0.000	0.000
AVEMP1	181	721	27.37	29.61	0.007	0.694
CE11	188	716	12.79	11.51	0.029	0.096
COAD101	189	724	6.08	5.41	0.033	0.016
COAD107	185	724	6.92	6.75	0.018	0.630
LIM111	192	751	7.97	7.74	0.026	0.299
CUS12	50	231	0.10	0.21	0.069	0.029
CUS14	50	231	0.00	0.14	0.046	0.005
EXP1	172	728	110.17	102.54	0.045	0.876
GROWTH1	204	750	2.96	3.05	0.031	0.049
INNOV15	115	521	0.45	0.58	0.014	0.016
LIM104	160	678	3.19	3.80	0.045	0.072
LIM106	156	660	2.76	2.08	0.042	0.026
MISEMP1	208	757	0.87	0.95	0.000	0.001
OVRDFT11	193	753	0.80	0.85	0.064	0.086
PARTARR1	206	756	0.25	0.31	0.078	0.068
PARTN12	21	234	0.37	0.65	0.039	0.012
SALES13	187	745	27.26	31.13	0.096	0.182
SYSTRD13	144	595	0.35	0.43	0.070	0.065
TOTFP1	192	714	20.94	30.92	0.014	0.005
TURN1	163	710	890.85	1233.56	0.124	0.041

* Equal variances not assumed.

Table 5. A comparison for cases where ICC employment data is present of CBR survey variables for companies grouped by missing/non missing CBR employment data.

	Numbers		Mean		Significance Level	
CBR Variable	Missing	Non Missing	Missing	Non Missing	Mann Whitney 'u'	Students 't'*
ADDFIN12	7	111	0.00	2.25	0.661	0.096
ADDFIN14	7	111	0.00	3.80	0.530	0.017
ADDFIN17	7	111	0.00	2.19	0.033	0.007
CE14	10	188	42.00	46.43	0.026	0.110
ADVICE11	10	195	0.00	0.00	0.347	0.000
COAD104	10	186	8.20	6.73	0.069	0.022
COAD103	10	188	7.70	6.51	0.245	0.082
CUS16	2	71	0.50	0.00	0.005	0.528
EXP1	3	191	0.00	90.70	0.294	0.011
MISPROF1	10	193	0.89	0.76	0.074	0.204
PARTN11	2	73	0.50	0.15	0.043	0.385
PARTN12	2	73	0.50	0.47	0.093	0.419
PROC1a	6	141	0.20	0.55	0.061	0.000
RD14	6	110	0.00	0.52	0.270	0.000
SALES13	8	186	0.00	0.30	0.220	0.000
SALES14	8	186	18.25	33.62	0.003	0.380
SYSTRD13	7	157	0.71	0.38	0.073	0.119

* Equal variances not assumed.

Table 6. A comparison for cases where ICC profits data is present of CBR survey variables for companies grouped by missing/non missing CBR profits data.

	Numbers		Mean		Significance Level	
CBR Variable	Missing	Non Missing	Missing	Non Missing	Mann Whitney 'u'	Students 't'*
ADDFIN15	33	113	0.30	3.08	0.252	0.042
ADDFIN17	33	113	0.30	2.95	0.167	0.006
ADVICE13	50	195	0.14	0.28	0.40	0.018
CEAGE1	46	184	49.26	46.15	0.050	0.071
COAD101	44	184	6.25	4.96	0.270	0.008
COAD103	46	191	7.00	6.35	0.913	0.087
CUS16	15	71	0.13	0.25	0.349	0.045
FINANCE1	46	194	0.74	0.61	0.111	0.095
LIM105	38	173	3.00	4.66	0.022	0.012
INNOV15	28	132	0.32	0.49	0.100	0.095
MISEMP1	50	195	0.86	0.96	0.005	0.047
RD14	28	115	1.96	1.86	0.042	0.210
SALES13	41	190	20.66	35.31	0.022	0.012

* Equal variances not assumed.

Table 7. Matching CBR and ICC Employment Data.

Parameter	Estimate	Std. Error
α_1	3.277	(1.203)
α_2	10.823	(1.213)
α_3	30.153	(2.446)
α_4	73.755	(3.619)
IccEmp	0.003	(0.076)
IccEmp^2	0.004	(0.0004)

$N = 195$
$R^2 = 0.867$

Table 8. Descriptive Statistics for Icc91 Profit Data.

Variable	mean	std.	skew	kurt	min	max	N
CBR Profits	68.5	185.4	5.0	40.7	-482.0	1738.0	181
				All Firms			
	-6.6	169.1	-6.3	62.0	-1783.0	649.0	243
ICC Profits							
				Matching ICC CBR Firms			
	-3.1	175.5	-6.6	66.1	-1783.0	649.0	181

Table 9. Matching CBR and Profit Data.

Parameter	Estimate	Std. Error
α_1 $(n = 1)$	-572.19	428.46
α_2 $(n = 4)$	-245.46	120.45
α_3 $(n = 125)$	25.154	15.188
α_4 $(n = 18)$	120.17	34.85
α_5 $(n = 90)$	55.289	16.909
α_6 $(n = 1)$	49.121	138.75
α_7 $(n = 8)$	-576.42	335.37
IccProfit	1.032	0.268
IccProfit^2	0.003	0.001
IccProfit^3	$1.3e-006$	$3.5e-07$

$N = 180$
$R^2 = 0.600$

Table 10. Regression-Based Imputation For Profit Data: *MAR* Assumed

Parameter	Estimate	Std. Error
Constant	155.09	119.22
Turn	−0.0138	0.014
Ceage	−24.248	0.063
Large	1.1835	0.837
Comps1	6.430	39.966
Skill	53.138	39.867
Train	−16.403	5.417
Limab101	0.088	2.211

$T = 105$
$R^2 = 0.346$

Table 11. Probability of Missing Data Model: Profits

	Parameter	t-ratio	Parameter	t-ratio
Constant	-4.146	-4.129	-4.019	-4.123
turn1	$-3.536e-005$	-0.319	$-4.92e-005$	-0.443
exp1	0.000	1.294	0.000	1.241
large1	0.162	1.499	0.136	1.293
comps11	0.000	0.086	0.002	0.041
skill1	0.365	1.211	0.334	1.182
train1	0.058	0.184	0.025	0.084
ceage1	0.026	1.818	0.085	1.820
TotMiss91	0.361	3.384	0.357	3.354
Finance1	0.959	2.424	1.031	2.671
ICCP91	-0.0021	-1.944	–	–

$T = 105$
$LogL = -543.69$

Table 12. Regression-Based Imputation for Profits Data:
MAR Not Assumed

Parameter	Estimate	Std. Error
Constant	−398.51	279.15
Turn	−0.021	0.015
Exp	0.389	0.066
Ceage	−9.511	15.737
Large	1.373	0.830
Comps1	77.736	51.457
Skill	56.960	39.548
Train	3.7169	2.738
Limab101	−14.588	5.4075
corrF	177.12	80.979

$T = 105$
$R^2 = 0.377$

Table 13. A Comparison of Imputation Methods: Profit Data

Comparisons			
Hot Deck	$\Sigma_{IMP}/\Sigma_{OBS}$	MABS	Mean
I Train			
Adj	0.223	84.22	-22.26
II Train and Growth			
Adj	1.365	86.41	10.46
III Growth and Train			
Adj	1.026	87.53	0.750
IV Exports, Finance and Train			
Adj	1.689	97.05	19.766
Regression			
Assuming Missing at Random			
Adj	−50.30	240.44	
Do Not Assume Missing at Random			
Adj	−0.568	81.838	

Table 14. Multiple Imputation (Mean): Profit Data

			Σ Im$p/\Sigma OBS$			
		Deciles			quinquartiles	
	10	100	1000	10	100	1000
Adj	0.793	0.733	0.773	0.700	0.992	0.998
			MABS			
		Deciles			quinquartiles	
	10	100	1000	10	100	1000
Adj	28.81	21.22	22.56	88.7	101.12	37.09
			Mean			
		Deciles			quinquartiles	
	10	100	1000	10	100	100
Adj	-9.61	-12.41	-10.54	-13.93	-0.368	-0.066

Figure 1: Patterns of Missing Data: Univariate Non-Response

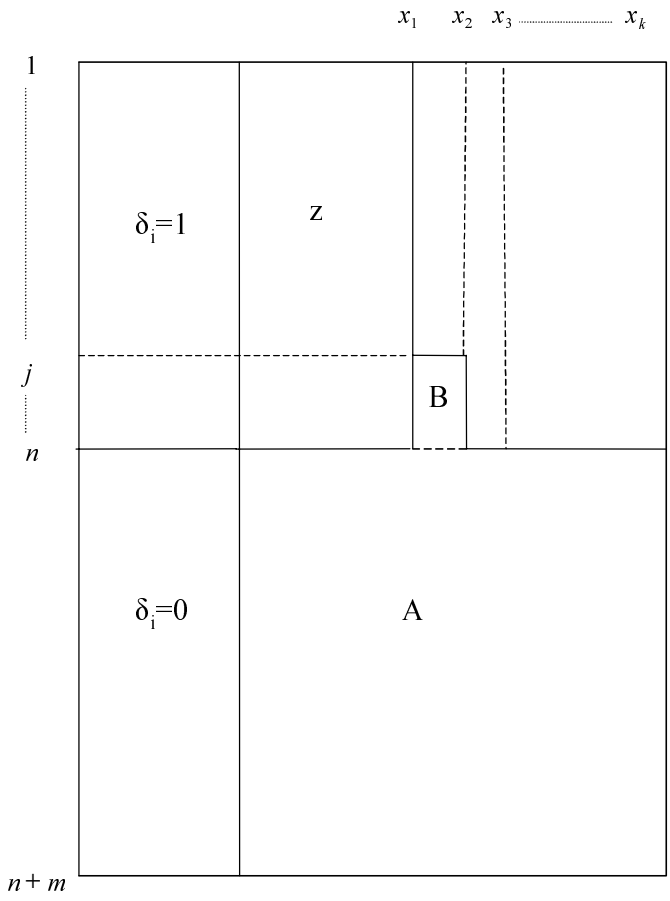

Figure 2: Patterns of Missing Data: Multivariate Non-Response

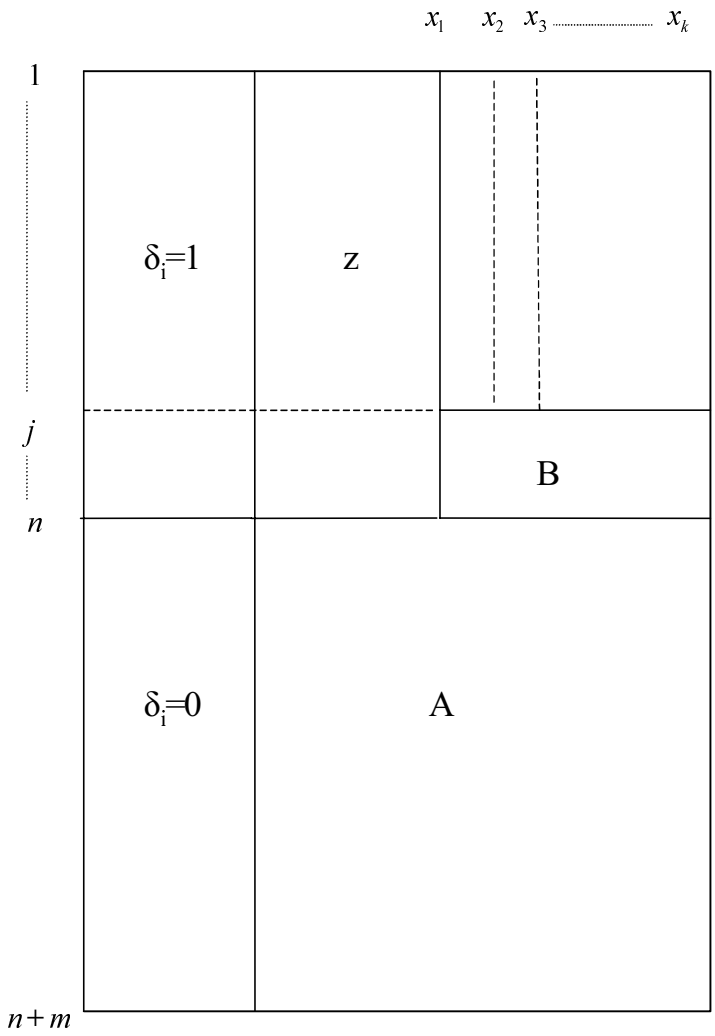

Figure 3: Pattern of Missing Data: Monotone Non-Response