Imputation of Missing Values in Survey Data: An Experimental Approach

End of Award Report

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

April 5, 2001

1. Background

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 and government sponsored projects involve the creation of complex datasets, many of which are based upon sample survey of firms or individuals. This is true, for instance, of the major surveys of innovation activity in the EU countries carried out under the Community Harmonised Innovation Survey Programme. 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 related issues such as selection bias are paramount.

In the case of the ESRC funded CBR small and medium sized business survey and other business based surveys, the problem of missing values can manifest itself in a number of important ways, for instance in a relatively greater reluctance to report profits data compared to categorised data such as questions on growth objectives. This problem is then manifest in the problem of both *item* (or question specific) non-response, and in a tendency for "better" firms to respond generally leading to *unit* (or questionnaire) non-response.[1]

The consequences of missing data will vary according to a large number of factors. These include:

(i) *objectives of the analysis*

(ii) *the pattern of missing data.* Critical here is whether missing data is univariate or multivariate. If, for example, the problem of missing data in a business survey affects a large number of variables, and that the likelihood

[1] For a particular discussion in relation to missing innovation item response rates see Cosh and Wood (1999)

of firms responding differs according to the question asked, then the set of firms for which there exists a complete set of responses may be quite small.

(iii) *the ratio of missing to non-missing cases.* For any single variable the consequence of non-response will depend, in part, upon the frequency of non-response. In the case of univariate analysis, and assuming that non-response is a random phenomenon[2], missing data will simply result in an efficiency loss, thereby compromising the ability to discriminate between alternative hypotheses. In a multivariate setting, the existence of missing data in one variable will often generate missing values in others; consider the case, for example, of the computation of bivariate correlation coefficients which can only be computed across the set of cases for which the two variables are both non-missing. If we move from a bivariate to a multivariate setting, then the pattern of missing data across a set of variables can result in a substantial loss of information. For example, some software packages will only perform statistical analysis on *rectangular* datasets, namely a $n \times k$ matrix, say M, of n firms with fully observed data over k variables. Again, even if the data in M is a representative sample of the total data (including firms with both complete and partially observed records over k variables), the efficiency loss can be substantial, and may result in statistically insignificant effects masking otherwise strong relationships.

(iv) *the missing data generating process.* Although we do not subscribe to the sometimes held notion that the search for a good economic model is a search for what is often referred to as underlying *data generating process*, (*dgp*) it is instructive to entertain the idea that model selection should seek to approximate the true and unknown *dgp*. In the same spirit, and faced with the problem of missing data, we can productively think of a missing data generating process (*mdgp)* which, in the form of a probability model, determines the likelihood of whether or not information on a given question is supplied. If for example, an analyst were to randomly discard the profit responses for a set of firms then the *mdgp* would be ignorable, in the sense that subse-

[2] See Weeks (2001) for a thorough overview of random and non-random non-response.

quent inference could proceed using only the observed set of responses; the only loss would be the precision in which model parameters were estimated. Alternatively, if the *mgdp* were such that firms only reported profits if profits were less than some threshold value then there would be both a loss of efficiency and bias.

At the outset we emphasise the importance of distinguishing between database providers and end-users. As noted above, the consequences of missing data in surveys will depend in part upon the objectives of the analyst. However, it is also true that there is likely to be substantial heterogeneity across the users of the data with respect to the ability to both select and handle missing data. This knowledge is in part subject specific. For example, users with substantial training in statistics have, since the seminal work of Rubin (1977), realised that imputed data should be treated differently from observed data in subsequent statistical analysis. In contrast some applied economists have followed a classical approach by providing *single* estimates of missing values without making allowance for an overall increase in uncertainty attached to these values. This issue is further discussed in Section 3 of this report.

2. Objectives

The principal objectives of this research follow naturally from a quote from Horowitz and Manski (1998).

> *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)

Faced with a number of competing methodologies with which to impute missing values we adopt an experimental approach by utilising a secondary data source

which facilitates evaluation. This data exists in the form of the ICC (Inter Company Comparison) database which holds the standardised financial accounts on the returns submitted by companies to Companies House as part of their statutory reporting requirements. Thus, once we have matched firms in the CBR database with ICC data, we will utilise the observed profit and employment data in ICC to evaluate the accuracy of imputation of missing values in the CBR database.

In the original proposal we outlined the following objectives:

i) construct a probabilistic model of non-response for profit and employment data;

ii) evaluate the performance of alternate techniques for imputing missing profit data.

iii) document and make available user-friendly routines for implementing various imputation techniques.

Within the confines of this one year long project, each of these objectives have been met and, we believe, that the outcomes of the research go beyond those stated.[3] In the case of i) Weeks and Hughes (2001) in Section 4 and demonstrate how the probabilistic model of non-response for both profits and employment data is integral to two competing imputation methodologies: a) a conditional mean model to predict missing values with a correction for non-random non-response; and b) a multiple imputation approach to imputing missing values. In addition, the probabilistic model of missing data, essentially a binary probit model using observed firm characteristics as predictors of missing data, is also used to test for whether the process generating the data is some function of the level of the variable in question. Note that by adopting an experimental approach and by using matched observations in the ICC database, we are able to do this. Further details are provided in Section 4 of Weeks and Hughes (2001).

The evaluation of the performance of alternative techniques (point ii) in the original objectives) has been carried out in Weeks and Hughes (2001) (see Section

[3] An important caveat is that to date we have focussed upon missing profits data. For reasons elaborated upon in Weeks and Hughes (2001), there were very few firms with both missing *employment* data and *observed* ICC data.

8.1). Although we believe that we have been successful in meeting this objective, there were a number of particular problems that require emphasis. First, in the case of profits data there were unanticipated problems in reconciling data for firms with reported profits data in *both* the CBR and ICC databases. The ICC and CBR profit data show significant differences. Given that we intend to use the observed ICC profits data (for firms not reporting profits in the CBR database) to evaluate predicted values for different imputation techniques, we obviously need data which is comparable. However, as discussed in Weeks and Hughes (2001), it is not necessary to have a perfect match in the case of *ranking* performance across different approaches.

With regard to objective (iii) the computations were completed using SOLAS for missing data and the Ox (see Doornik (1999)) programming language. Although SOLAS includes a module to perform multiple imputations[4], this proved quite inflexible and difficult to use. As a result we developed our own module, the code for which is provided in Appendix 1 of Weeks and Hughes (2001).

We believe that we have surpassed our original objectives in the following sense. First, the literature review on missing data completed by Weeks (2001) represents a comprehensive overview of the general problem of missing data, providing a taxonomy of approaches which also includes a number of new and innovative techniques such as generalised entropy. Second, in completing the work we now feel able to advise other database providers on the relative merits of imputation strategies.[5] This, we believe, represents a significant outcome of the research. Further fruitful work on modelling nonresponse in business profits data could be developed on the foundations laid in this short project by developing a more sophisticated behavioural model of business profits.

3. Methods

The methods used in the report are, with the exception of one of the imputation methods, very familiar to researchers across the social sciences. These methods,

[4] See SOLAS (1997).

[5] The Department of Trade and Industry have asked us to present a seminar on imputation techniques. See Section 7.

essentially regression-based tools, are used as inputs into relatively straightforward techniques for imputing missing values. For example, consider the case where we observe profit data for n_1 firms, with n_0 non-respondents. In order to *forecast* profit information for the n_0 non-respondents we might specify a model of profits for respondents writing,

$$P_i = \alpha + \boldsymbol{\beta}\mathbf{x}_i + \varepsilon_i, \quad i = 1, ..., n, \tag{3.1}$$

where i indexes the set of firms with observed profits, $\boldsymbol{\beta}$ is a $1 \times k$ vector of unknown parameters, and $\mathbf{x}_i$ is a $k \times 1$ vector of variables which we believe in combination provide predictive information for P_i. Given an estimate of $\boldsymbol{\beta}$, say $\widehat{\boldsymbol{\beta}}$, and fully observed data in $\mathbf{x}$, we may then impute missing profit data using the simple relationship, $\widehat{P}_j = \widehat{\alpha} + \widehat{\boldsymbol{\beta}}\mathbf{x}_j$, where $j = 1, ..., n_0$ indexes non-respondent firms. This method represents one of the basic techniques for imputing missing data, and depends critically on the assumption that the process determining whether or not firms supply profit information is unrelated to the level of profits. If, for example, only firms with profits exceeding some threshold value, ζ, responded to the profit question, then the use of a set of parameters $\widehat{\boldsymbol{\beta}}$ to impute data would result in biased imputations. Both Weeks (2001) and Weeks and Hughes (2001) elaborate upon this method, also detailing a procedure which provides a correction in the case where the missing at random assumption is invalid.

The two methods which we use which might not be so familiar to all social scientists are the Hot Deck (HD) and Multiple Imputation approach to imputing missing data. Both these methods, although involving some relatively sophisticated details, are very intuitive and easy to motivate. HD, which has been a standard approach to forecasting missing data in the US Censuses, is based upon the fundamental notion of *matching*. That is, if for a given firm we do not observe profit data, it makes sense to use *donor* values from other similar firms for which data is reported. For example, if in the unlikely case we believe that the distribution of profits is bi-modal,[6] and that there exists an almost deterministic relationship of the form

$$\textit{Firms that train} \quad \equiv \quad \textit{High profits} \tag{3.2}$$

[6] Bi-modal with the variance around each mode being negligible.

$$\textit{Firms not train} \quad \equiv \quad \textit{Low profits}$$

then we could reasonably sample from the subset of firms with observed profits and which train their employees, and use the profit information as donor values for non-respondent firms which train. Obviously as soon as we depart from this purely pedagogical case, we need to control for other factors, which in combination with training determine the level of profits. However, if the relationship between training and profits is not as we have depicted then one way to account for the subsequent distribution of profits within each of the four cells[7] is to sample with replacement from each of the cells and compute an average.

One of the principal distinctions between the HD approach to imputation and the regression-based approach outlined above is that the HD approach is non-parametric in that rather than using the information provided by a parametric model, it utilises the principle of matching. A variant of the HD approach which might be considered as a derivation of a Bayesian approach to inference, is the Multiple Imputation technique. This approach is predicated upon the specification of a parametric model of the missing data, with the rudiments of the methodology being as follows:

1. Specify a model of the missing data processes, say $\Pr(Mis_i|Obs_i)$, where Mis_i is equal to 1 if profit data is missing for firm i, 0 otherwise. Obs_i represents the observed information and would generally take the form of a vector of data for each firm, $i = 1, ..., n_1 + n_0$.

2. $\Pr(Mis_i|Obs_i)$ might take the form of a simple binary probit or logit model. Using the vector of estimated probabilities ($\widehat{P}_i$) representing, for each firm, an estimate of the probability that profit data is missing, $\widehat{P}_i$ is then partitioned into equal sized groups using quantiles.

3. Within each quantile there will be both firms with missing and those with observed profit data. In the same spirit as the HD approach, Multiple

[7] The four cells being: i) train; high profits, ii) train low profits; iii) not train, high profits; not train, low profits. Note that by referring to a *distribution* of profits in each cell we mean that the relationship is not deterministic as implied by (3.2).

Imputation proceeds by sampling (with replacement) from each quantile donor profits records and using this to fill in the missing values.

4. At this stage the way in which complete datasets are constructed may differ according to the analytical persuasion of the analyst. As discussed in Weeks and Hughes (2001), a classically trained analyst might resample many times and, for each non-respondent firm, and report the average value. In contrast, a Bayesian approach[8] might avoid reducing the information obtained in the multiple draws to a single average, and simply report multiple datasets, the multiplicity reflecting the additional uncertainty given that within the new complete database one or more of the variables contains imputed (i.e. predicted) information. This last point requires particular emphasis given that it is obviously erroneous to treat two variables, one of which for example, contains zero cases of missing data, the other exhibiting a relatively high incidence of missing values, on the same footing. A priori, there is far more uncertainty attached to the latter.

4. Results

This research project has conducted an evaluation of a number of competing imputation techniques for missing data in sample surveys. Applied to the particular problem of missing profit data in the Centre for Business Research (CBR) Small and Medium Sized Enterprise database, the defining characteristic of this research is the availability of a secondary data source, providing benchmark data for non-respondent firms. This data is derived from the Inter Company Comparison (ICC) database, containing data which firms must deposit at Companies House.

This particular research was motivated by the problem of determining which of a number of competing approaches was best in terms of imputing missing data. Obviously, the general problem is that most analysts are confronted with the situation that any assumptions that are made as to the pattern of missing data are not verifiable. For example, simple regression-based procedures are frequently

[8]For example, the SOLAS for Missing Data Statistical Software does not facilitate easy calculations of these averages, choosing to reported multiple datasets (see SOLAS (1997)).

used on the basis that the process generating the missing data is independent of the level of profits. Such a situation would arise if, for example, a database manager randomly discarded the profit responses for a group of firms. With access to a secondary source these type of assumptions can be tested, thereby facilitating a more informed choice of imputation technique. In this study we found that the process of missing profit data was not random, and that the probability of reporting profits fell with the level of profits.

The techniques evaluated range from simple regression-based methods with and without corrections for non-random nonresponse, to matching procedures designed to locate firms which, if similar on a set of observed firm-level attributes, may (given a number of assumptions) can be used as *donors* i.e. supplying profit data for otherwise similar firms which do not report profits. The principle problem encountered in undertaking the research was the difficulty in reconciling the two sources of data. This is obviously important given that we use the ICC data to evaluate the performance across the different imputation techniques. Thus, in interpreting our results we recognise that imputation error will contain two components: a constant term (across techniques) reflecting the lack of comparability, and true prediction error which will vary across techniques. Despite this problem we are able to identify a definite ranking across the techniques: the resampling technique, referred to as Multiple Imputation performs best, with both regression-based and matching procedures providing disappointing results.[9] An important caveat on these findings is added below.

In identifying directions for future research we emphasise that our analysis has been conducted in a univariate setting. That is we have conducted imputation and evaluated different techniques by focussing upon the incidence of missing data in a single variable. Although we can readily motivate the importance of profit data, and why one might expect systematic patterns in nonresponse, it is also the case that in large databases the incidence of missing data is widespread, and importantly, the pattern of missing data is correlated across survey questions. We found this to be the case in this study, noting the explanatory power of a variable capturing the extent of missing data across a wide set of firm characteristics in a

[9] See Weeks and Hughes (2001) for details.

model of *missingness* for profits. Subsequently we believe that a useful extension of the research conducted here would be to work with *multivariate* models of missing data.

5. Activities

Weeks is an Associate of Cambridge Econometrics, a local firm which provides consultation advice for, among others, the European Commission. During 1998 he was involved in a project for Eurostat, the Statistical division of the Commission, on approaches to imputation in the context of missing data within CRONOS, the pan European database on regional employment, industrial output and growth.[10] Cambridge Econometrics have expressed interest in the findings of this research and many utilise some of the methods used in database management.

In April 2000 Weeks attended a conference on methodological issues in the imputation of missing data at the Wellcome Genome Campus, Cambridge[11]. Primarily sponsored by the software company Statistical Solutions, the conference included a number of prominent statisticians including Professor D. Rubin from Harvard University.

Hughes is a member of the ONS/DTI study group on the utilisation of the Harmonised Community Innovation Datasets and is responsible for the conduct and analysis of the biennial CBR SME Survey. An analysis of response rate bias utilities was published in Cosh and Hughes (2000). Hughes and Weeks will present a seminar on imputation methods to the DTI and Cambridge Econometrics. (See Section 7).

6. Outputs

There are two principle publications based upon the missing data project.

a. *Methods of Imputation for Missing Data*: Melvyn Weeks

[10] See Weeks (1998).

[11] *Challenging Statistical Issues in Clinical Trials*, April 20th, 2000, Wellcome Genome Campus, Cambridge.

b. *Missing Observations in Survey Data: An Experimental Approach to Imputation*; Melvyn Weeks and Alan Hughes.

The first paper reviews the considerable literature on missing data and provides a taxonomy of missing data problems. Of particular note is the discussion of entropy-based methods of imputation which are potentially attractive given the minimal use of assumptions as to the structure of the data. In b) we apply a number of imputation techniques to the problem of missing data in the Centre for Business Research database on small to medium-sized companies.

7. Impacts

As a result of the work of Hughes with DTI in the design and piloting of the CIS3 Innovation Survey, Hughes and Weeks have been invited to present a workshop to be organised by the DTI Survey team on Methods of Imputing Missing Values in the forthcoming UK component of CIS3 Survey.

8. Future Research Priorities

Our future research priorities relate directly to the problems we have encountered in undertaking this research. First, we aim to create a dataset which has a larger number of firms with both *missing* data for the CBR variable and *observed* secondary data - in this case the ICC dataset. Second, we believe that a useful extension of this research would be to conduct the analysis in a multivariate setting. To date we have focussed upon the problem of missing profits data, but believe that there would be substantial gains to model the process of *missingness* within a multivariate framework.

References

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.

Cosh, A. D., H. A., and E. Wood (1999): "Innovation in UK SMEs: Causes and Consequences for Firm Failure and Acquisition," in *Entrepreneurship SMEs and the Macro Economy*, ed. by Z. Acs, B. Carlsson, and C. Karlsson. Cambridge University Press, Cambridge.

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

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.

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.

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.

SOLAS (1997): *SOLAS for Missing Data Analysis 1.0*.Statistical Solutions Ltd., 8 South Bank, Crosse's Green, Cork, Ireland.

Weeks, M. (1998): "Methods of Imputation for Missing Data," Report for Eurostat relating to the project Model Based Regional Indicators.

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

Weeks, M., and A. Hughes (2001): "Missing Observations in Survey Data: An Experimental Approach to Imputation," mimeo, Centre for Business Research, University of Cambridge.