DataFirst Technical Papers

Univariate Multiple Imputation for Coarse Employee Income Data

by

Reza C. Daniels

Technical Paper Series

Number 17

About the Author(s) and Acknowledgments

Reza C Daniels - School of Economics and SALDRU, University of Cape Town, reza.daniels@uct.ac.za

This document has been produced as a chapter of my doctoral dissertation. I am grateful to the African Economic Research Consortium and the National Research Foundation for funding. I am also grateful to my supervisers – Murray Leibbrandt and Martin Wittenberg – for their guidance. All errors and/or omissions are the responsibility of the author.

This is a joint DataFirst and SALDRU paper.

Recommended citation

Daniels, R.C. (2012). Univariate Multiple Imputation for Coarse Employee Income Data. A DataFirst Technical Paper Number 17. Cape Town: DataFirst, University of Cape Town

DataFirst, University of Cape Town, Private Bag, Rondebosch, 7701, Tel: (021) 650 5708,
Email: info@data1st.org/support@data1st.org

Univariate Multiple Imputation for Coarse Employee Income Data[*]

Reza C. Daniels[†]

Abstract

This paper is concerned with conducting univariate multiple imputation for employee income data that is comprised of continuously distributed observations, observations that are bounded by consecutive income brackets, and observations that are missing. A variable with this mixture of data types is a form of coarsening in the data. An interval-censored regression imputation procedure is utilised to generate plausible draws for the bounded and nonresponse subsets of income. We test the sensitivity of results to mis-specification in the prediction equations of the imputation algorithm, and we test the stability of the results as the number of imputations increase from two to five to twenty. We find that for missing data, imputed draws are very different for respondents who state that they don't know their income compared to those who refuse. The upper tail of the income distribution is most sensitive to mis-specification in the imputation algorithm, and we discuss how best to conduct multiple imputation to take this into account. Lastly, stability in parameter estimates of the income distribution is achieved with as little as two multiple imputations, due largely to (a) the small fraction of missing data, in combination with (b) reduced within- and between-imputation components of variance for imputed draws of the bracketed income subset, a function of the defined lower and upper bounds of the brackets that restrict the range of plausibility for imputed draws.

Key Words: Multiple Imputation, Coarse Data, Income Distribution

JEL Codes: C15, C83, D31

[*] This document has been produced as a chapter of my doctoral dissertation. I am grateful to the African Economic Research Consortium and the National Research Foundation in South Africa for funding. I am also grateful to my supervisers – Murray Leibbrandt and Martin Wittenberg – for their guidance, as well as three external examiners for their comments. All errors and/or omissions are the responsibility of the author.

[†] School of Economics & SALDRU, University of Cape Town. reza.daniels@uct.ac.za

1 Introduction

Employee income data are often coarsened as a result of questionnaire design. Statistics South Africa (SSA), for example, generally ask two sequential employment income questions: an exact income question with a showcard follow-up (see Daniels, 2012 for further discussion). In public-use datasets, this results in two income variables: a continuously distributed variable for exact income responses and a categorical variable for bounded income responses with separate categories for nonresponse. It is the task of the researcher to then generate a single income variable that effectively deals with this mixture of data types. Following Heitjan and Rubin (1991), we call a variable with this mixture of data types "coarse data".

Coarse income data pose non-trivial implications for researchers concerned with analysing that data. The primary problem that arises from an inconsistent treatment of this variable is that parameter estimates may be biased and dependent on the particular researcher's choice of method to overcome the problems posed by the instrument's design and resulting data structure. This leads to potentially erroneous inferences on important univariate parameters of the income distribution, including quantiles and moments.

Multiple imputation is potentially an effective solution for coarse data problems (Heitjan and Rubin, 1990; Heitjan, 1994). It involves substituting coarse data values with plausible draws of those values multiple times. Multiple imputation has been applied to coarse wealth data by Heeringa (1995) and Heeringa, Little and Raghunathan (2002), and it has been applied to coarse earnings data by Daniels (2008) and Vermaak (2010). Ardington, Lam, Leibbrandt and Welch (2006) conducted multiple imputation for total income. However, because multiple imputation is effectively a simulation-based technique (Schafer, 1999), it is very dependent on the setup of the imputation process and can frequently perform sub-optimally for reasons that may not be easy to isolate. Van Buuren, Boshuizen and Knook (1999), Royston (2004), White, Wood and Royston (2007) and Graham, Olchowski and Gilreath (2007) discuss various aspects of the multiple imputation process that can affect the reliability of imputed draws and statistical inference, ranging from covariate selection, the imputation algorithm itself and the numbers of imputations needed for reliable inference.

In this paper the imputation algorithm is simplified by imputing univariately for coarse income data only, rather than also imputing covariate missing data. This has both advantages and disadvantages. The main disadvantage is that it removes all units with covariate nonresponse from the estimation sample, which is equivalent to treating covariate nonresponse as missing completely at random (MCAR). The cost of doing this is dependent on the application, however, with Allison (2000) noting that more sophisticated treatments of covariate nonresponse can impose equally stringent (but often more opaque) assumptions on the data. However, a distinct advantage of multiple imputation is that imputed draws can be made for many variables with missing data simultaneously, making it computationally efficient. There is, therefore, a definite trade-off in ignoring covariate nonresponse.

The main advantage of imputing multiple times for a single variable is that it allows us to be far more precise about exactly which aspects of the multiple imputation algorithm lead to implausible results. The two primary dimensions of the imputation algorithm that will be explored are specification of the prediction equations and sensitivity of the results to the number of imputations. The reason we need this precision is because Daniels (2012) showed that respondents who chose to answer the bounded income question generally were higher income individuals. However, when we accounted for predictors of higher incomes in the sequential response propensity models, it was revealed that the final nonresponse subset had refusals that were largely indistinguishable from don't know responses on observable covariates. It was this finding that led to the suggestion that final nonresponse was likely an ignorable form of nonresponse.

A key objective post-imputation is then to assess where in the income distribution the bounded, refuse, don't know and unspecified subsets of the employment income question lie. The coarse data framework allows us to characterise the nature of the problem in a theoretically sound manner. The simplified univariate multiple imputation algorithm then allows us to test the sensitivity of inferences to covariate selection and the number of imputations. The usefulness of doing this is that we learn how robust imputations are to mis-specification. Lessons learnt from this process can then feed into more complex multivariate missing multiple imputation exercises.

In order to examine the performance of the imputation algorithm, we test four different specifications of the prediction equations: one that is completely mis-specified to establish a baseline of how wrong the imputed draws can be; one with covariates selected identically to the response propensity models of Daniels (2012); one with Mincerian earnings function based covariates; and one with a combination of response propensity and Mincerian earnings function covariates, which we treat as the first-best specification method for reasons discussed below.

Data for this exercise includes the October Household Surveys (OHS, 1997-1999) and Labour Force Surveys (LFS, 2000-2003 September Waves only). The sample is restricted to economically active (16-64 year old) employees only.

2 Preliminaries

2.1 Coarse Income Data

A variable with continuous, bounded and missing observations is not simply an example of nonresponse, but in fact a more complicated problem known in the literature as "coarse data". The theory of coarse data stems in part from the theory of missing data, which was principally developed by Rubin (1976, 1987). However, "coarse data" is in fact a generalisation of the various ways that data may not reflect their true values, and includes as special cases rounded, heaped, censored, partially categorised and missing (i.e. completely coarse) data (Heitjan and Rubin, 1991).

Two principal papers established the theory of coarse data: Heitjan and Rubin (1991) and Heitjan (1994). To show the direct precedents to missing data theory, it is useful to note that the theory of coarse data generalised Rubin's (1976, 1987) theoretical phraseology–an association partially mandated by the result that missing data was simply one form of coarsening. As a consequence, the concepts of missing completely at random" (MCAR), "missing at random" (MAR), and "not missing at random" (NMAR) were distinguished from "coarsened completely at random" (CCAR) and "coarsened at random" (CAR). Heitjan and Basu (1996) explicitly differentiate between these five concepts, but the epistemological extensions provided by coarse data theory are particularly useful to income in public-use micro datasets.

For the purposes of this discussion, coarse data is defined to consist of a combination of continuous data (assumed not to be coarsened at all), bounded data (bracket responses), and item missing data. We formally define what this means for the univariate statistical distribution of income, commencing with the missing data framework and then incorporating the more general coarse data framework.

Following Little and Rubin (2002, 12), we define the complete data matrix as $Y = (y_{ij})$ and the missing data indicator matrix $M = (M_{ij})$. Y is differentiated into an observed and unobserved component, Y_{obs} and Y_{mis}. The distribution $f(\cdot)$ of missingness is conditional upon Y and unknown parameters ϕ, denoted $f(M|Y, \phi)$. If $f(M|Y, \phi) = f(M|\phi) \; \forall \; Y, \phi$, the unobserved data are said to be Missing Completely at Random (MCAR). Here, missing data do not depend on the observed or unobserved components of the complete data matrix. If $f(M|Y, \phi) = f(M|Y_{obs}, \phi) \; \forall \; Y_{mis}, \phi$, the unobserved data are said to be Missing at Random (MAR), a more restrictive condition than MCAR because now the missing data depend on the observed data. If the missing data M depend on the missing values in the data matrix, the mechanism is called not missing at random (NMAR). The missing data mechanism is said to be "ignorable" if the unobserved data are thought to be MCAR or MAR; in this case, a separate model for the mechanism that causes non-response is not needed (i.e. can be ignored). The missing mechanism is said to be "non-ignorable" if the unobserved data are NMAR.

The coarse data framework incorporates missing data as a type of coarsening, but is also generalisable to bounded data such as income reported in brackets. To see the extensions, we again rely on Little and Rubin's (2002, 127-129) formulation of the problem. Let Y be the complete data matrix in the absence of coarsening with sample space Ψ, and let $f(Y|\phi)$ denote the density of Y for the complete data with unknown parameters ϕ. The observed data are now thought to consist of a subset of the sample space Ψ in which Y is known to fall. This subset is a function of Y and a coarsening variable G that determines the bounds of Y_{obs}, so that $Y_{obs} = Y_{obs}(Y, G)$.

To see the extension to bracketed responses such as those present in income microdata, note that the characterisation of $Y_{obs} = Y_{obs}(Y, G)$ assumes that the observed data fall within *known* upper and lower bounds and not outside these bounds. Since the bounds are assumed known, the coarse data framework is flexible enough to be applied not only to bracketed response types, but also to data that is thought to be imprecisely coarsened, such as rounded data, heaped data, or otherwise partially categorised data (see Heitjan and Rubin, 1991). In each case the coarsening mechanism needs to be precisely modelled.

To incorporate missing data into this framework, call the unobserved data completely coarsened, and allow plausible values of that data to lie within the sample space Ψ of Y. In this case, G is simply the missing data indicator matrix. Thus:

$$y_{obs,ij} = \begin{cases} \{y_{ij}\}, & \text{the set consisting of the single true value, if } G_{ij} = 0 \\ \Psi, & \text{the sample space of Y, if } G_{ij} = 1 \end{cases} \tag{1}$$

From this, the data Y_{obs} are called coarsened at random (CAR) if $f(g|y_{obs}, y_{mis}, \phi) = f(g|y_{obs}, \phi)$ for all y_{mis}.

To apply the framework to a mixture of continuous responses, bounded responses and missing data, we follow Heeringa's (1995) example and simply allow G to precisely define whether the data are observed as continuous, bracketed or missing. To make the framework specific to the income question in the OHS and LFS, we will characterise the coarsening process to match what is found in the public-use datasets.

$$y_{obs,ij} = \begin{cases} \{y_{ij}\}, & \text{if } G_{ij} = \{0\} \\ [y_L \leq y_{ij} < y_U), & \text{if } G_{ij} = \{1, 2, ..., 14\} \\ \Psi, & \text{if } G_{ij} = \{15, 16, 17\} \end{cases} \tag{2}$$

Here, $G_{ij} = \{0\}$ indicates that y_{ij} is observed as a set consisting of the single true (exact) income value; $G_{ij} = \{1, 2, ..., 14\}$ indicates that y_{ij} falls within the lower bound y_L and upper bound y_U of one of the fourteen possible brackets in the OHS and LFS income questions; and $G_{ij} = \{15, 16, 17\}$ indicates that y_{ij} is observed as "Don't Know", "Refuse" or "Unspecified", and would then fall within the sample space of Y.

A key implication of the coarse data framework is that the variable G itself is measurement error free (Heitjan and Rubin, 1991; Wittenberg, 2008). This effectively implies that if a respondent reports their income to be within a given bracket, it cannot lie outside of those bounds. It also implies that if a respondent provides an exact income response, that response is assumed to be precisely reported. One of the implications of this relates to the imputation process for it implies that plausible draws of income for the bracketed subset of observations have to lie within the lower and upper bounds of those brackets, while draws for the missing data can be made over the sample space of income.

2.1.1 The Special Case of Unspecified Responses in the Coarse Data Framework

In Statistics SA's household surveys between 1997 and 2003, nonresponse to the employee income question was often recorded in the public-use data as an unspecified response. This response type exists even when there are options for don't know and refuse in the questionnaires. In 1999, the don't know option was introduced to the question for the first time, before both don't know and refuse options were added in 2000. Despite this, in each of the LFS, unspecified responses still exist for the subsample of employed economically active individuals. This represents a form of either processing or measurement error because don't know and refuse exhaust the possible nonresponse types in the income instrument.

Because of this, the nature of the coarsening mechanism for unspecified responses is opaque. Unspecified responses in the OHS 1997 and 1998 are the only identifiable form of nonresponse because the income question does not present any options to the interviewer for recording a don't know or refuse response. Therefore, we are forced to treat those as nonresponse. In 1999, the unspecified responses are confounded with refuse responses. But in the LFS, unspecified responses are identifiable as a form of processing error.

Observations that are deemed to be a result of processing error cannot simply be included in the coarse data framework as applied here, for it represents a mutually exclusive error mechanism in the data. We deal with this below by firstly exploring the extent of processing error in the data and then conducting independent multiple imputations for these observations.

2.1.2 The Special Case of Zero Income Brackets

An idiosyncratic feature of the bounded income question in all of the surveys analysed (OHS97-LFS03) is that it has a zero income option in the showcard. The existence of zero income brackets is thought to be related to false income reporting by Vermaak (2010), who imputes a proportion of these responses based on an assessment of the share that seem plausibly zero. The coarse data framework does not allow measurement error in the coarsening process to exist. Therefore, simply imputing the zero responses without a theoretical basis for doing so is arbitrary. Vermaak (2010) seems to include the self-employed in her subsamples of economically active individuals, which increases the number of zero responses substantially. This is easy to do in the LFS because the same question is asked to both the employed and the self-employed, whereas in the OHS the income question was different for self employed individuals. We restrict the sample here to employees only in all survey years.

Zero income values can exist as a valid response type for the subsample of economically active employees because respondents can be off work on unpaid leave. We evaluate the prevalence of zeros income responses below, but keep all such observations in the data without imputing them.

2.2 Multiple Imputation

Multiple imputation has gained recognition as one of the most effective methods for handling multivariate item nonresponse in public-use datasets. However, its use requires a clear understanding of its limitations. The coarse data framework is very useful for characterising the possible ways in which observed data may differ from their true values, and while it incorporates missing data as a type of coarsening, its extension to other data problems such as measurement error is limited on theoretical grounds. Recent advances in multiple imputation theory do indeed pose solutions to data measured with error (see, particularly, Ghosh-Dastidar and Shafer, 2003), but associated with this is (1) a necessary change in the operation of imputation algorithms and, (2) a modification of the combination rules required for valid statistical inference from multiply imputed datasets (Reiter and Raghunathan, 2007).

Multiple imputation has to address the pattern of coarsening present in a dataset. It was traditionally envisaged as a tool for data base constructors whose use of the methods was assumed to be independent from the data analyst's (Rubin, 1996). However, as the algorithms became more widely available and as more researchers became familiar with the methods, its use has burgeoned across the social and life sciences to a vast array of different applications. Indiscriminate use of multiple imputation is clearly discouraged by the major proponents of the method. As Schafer (1999) points out, multiple imputation is neither the only principled method for handling missing values, nor is it necessarily the best. Indeed, "(f)rom a statistical standpoint,

... a naive or unprincipled imputation method may create more problems than it solves, distorting estimates, standard errors and hypothesis tests..." (Schafer, 1999, 3). This view echoes Rubin's (1996: 475), who reminds all that the "actual objective (of multiple imputation) is valid statistical inference not optimal point prediction under some loss function, and replacing the former with the latter can lead one badly astray".

One of the important implications of the coarse data framework discussed in subsection 2.1, and directly implied by equation [2], is that the type of coarsening is defined to be precise; in other words, there can be no measurement error in the coarsening variable (G). The use of the coarse data framework thus places particular restrictions on the manner in which multiple imputation can be conducted. Its utility lies in the the fact that it provides clear rules for multiple imputation for the data structure resulting from the income question in the surveys considered.

There are examples in the literature of multiple imputation being used to deal with other forms of survey error. In particular, Ghosh-Dastidar and Shafer (2003) demonstrate how multiple imputation theory can be extended to the case of nonresponse and measurement error (without a validation study). They call their process multiple edit multiple imputation (MEMI), and note that producing MEMI's require assumptions about the distribution of the ideal data, the nature of nonresponse, and a model for the measurement error mechanism. This approach can also be adapted to suit other uses of multiple imputation, such as anonymising confidential survey information (ibid, 2003). However, in each case both the imputation algorithms and the rules for estimation and inference from the multiply imputed datasets differ, and have to be derived for the intended application.

3 Setup of the Problem

In this section we firstly discuss the data preparation tasks needed before working with the employee income variables. Here, the existence of bounded zero responses and processing error will be evaluated. We then develop an appropriate multiple imputation algorithm for coarse income data and identify the rules for estimation and inference given the nature of the coarse data problem and the imputation process.

3.1 Data Preparation

3.1.1 Zero Income Responses

Since the subsample of interest is economically active employees, zero income responses ought not to exist in general, unless the person is off work temporarily and on unpaid leave. However, in each survey year, there are a positive number of zero responses in the OHS and LFS. Moreover, the majority of zero responses are reported in the bounded income question in the OHS and LFS questionnaires, rather than the exact income question. Table 1 presents the number of observations reported in each response type.

Table 1: Distribution of Response Types: OHS97 - LFS03

Response Type		1997	1998	1999	2000	2001	2002	2003
Exact	Obs	16 185	7 637	11 735	18 739	15 945	14 469	13 759
	Percent	67.76	58.81	53.52	87.34	75.25	70.55	68.26
Exact-Zero	Obs	1	.	.	6	3	.	.
	Percent	0.00	.	.	0.03	0.01	.	.
Bounded	Obs	6 713	4 718	8 028	1 997	4 044	4 650	4 964
	Percent	28.10	36.33	36.61	9.31	19.09	22.67	24.63
Bounded-Zero	Obs	45	2	27	36	21	34	34
	Percent	0.19	0.02	0.12	0.17	0.10	0.17	0.17
Don't Know	Obs	.	.	1 588	72	521	651	485
	Percent	.	.	7.24	0.34	2.46	3.17	2.41
Refuse	Obs	.	.	.	144	578	664	891
	Percent	.	.	.	0.67	2.73	3.24	4.42
Unspecified	Obs	942	628	548	461	77	40	23
	Percent	3.94	4.84	2.50	2.15	0.36	0.20	0.11

Evident from the table is that the number of zero responses is usually very small, ranging from two in 1998 to forty-five in 1997. Most of these are reported in the bounded income question.

Of those employees who reported a zero income response (either in the bounded question or the exact question), the percentage that also reported that they have been absent from work in the past week due to illness ranges from zero in 1997-1999 to 29 percent in 2000, 42 percent in 2001, 53 percent in 2002 and 24 percent in 2003. There is no question for whether individuals are on unpaid leave for other reasons, however, so we cannot investigate this phenomenon. Because there are legitimate reasons for zero income reporting, we keep all zero responses in the subsamples of employees for each survey year and do not impute any of them.

3.1.2 Processing Error and/or Measurement Error in the Data

Two anomalies exist in Statistics SA's OHS and LFS: (1) instances where both an actual and a bracketed value are observed for the same individual; and (2) observations that are coded as "Unspecified" (i.e. missing), when in fact response options already exist in the questionnaire for the respondent to reply that they "Don't Know" or "Refuse" to answer the question. It is impossible to tell from the data or the survey documentation whether these anomalies are by design or whether they constitute a form of processing or measurement error, but they need to be addressed before imputation can taken place.

To formalise the problem, consider that the universe of potential outcomes for income responses consists of a continuous (exact) income subset, a bounded subset, and a missing (don't know, refuse or unspecified) subset. These three subsets are mutually exclusive because a bracketed outcome is only observed if the respondent chose *not* to answer the actual income prompt from the interviewer. A missing outcome is only observed if the respondent chose not to answer *both* the actual and the bracketed response prompt.

Let the event that an exact income response is reported by the respondent be denoted $P(A)$, the event that a bounded response is reported be denoted $P(B)$, and the event that a missing response be reported be denoted $P(M)$. For these three events to be mutually exclusive, $P(A \cup B \cup M) = P(A) + P(B) + P(M) = 1$, and $P(A \cap B \cap M) = 0$; $P(A \cap B) = 0$; $P(A \cap M) = 0$; $P(B \cap M) = 0$. A first form of (either processing or measurement) error can then be defined to exist if any of these outcomes are violated.

Because the design of the income question evolved between the OHS 1997 - LFS 2000, $P(M)$ is not defined by don't know and refuse for every survey year. We therefore need to decompose $P(M)$ into its observable parts: don't know responses (denoted $P(D)$), refusals (denoted $P(R)$), and unspecified responses (denoted $P(U)$). Across the survey years we will then observe missing responses as:

- $P(M) = P(U)$ for OHS 1997 and 1998;

- $P(M) = P(U) + P(D)$ for OHS 1999;
- $P(M) = P(D) + P(R)$ for LFS 2000-2003.

A second form of error can be defined to exist only for the LFS if $P(M) = P(D) + P(R) + P(U)$, where $P(U) \neq 0$. This is because don't know and refuse responses in the LFS complete the possible forms of nonresponse for the employed, economically active population. In the OHS 1999, unspecified responses cannot be identified as a form of error because those responses confound refusals in the same way that unspecified responses confounded both don't know and refusals in the OHS 1997 and 1998.

Table 2 presents the extent of these errors in the OHS97-LFS 2003. In order to estimate the subsets correctly, we use the raw data from the surveys of interest before any transformations of the variables are made.

Table 2: Subsets of Interest in the Observed Income Data

Income Response Subsets	1997	1998	1999	2000
N (employed EAP)	23 886	12 985	21 926	21 455
(1) Exact Responses: P(A)	0.6779	0.5881	0.5352	0.8737
(2) Bounded Responses: P(B)	1.0000	0.4888	0.8981	0.0951
(3) Nonresponse: P(M)	0.0000	0.0000	0.0724	0.0101
(4) Complement: $(A \cup B \cup M)^{\mathsf{C}}$	0.0000	0.0484	0.0250	0.0215
Sum: (1) + (2) + (3) + (4)	1.6779	1.1253	1.5307	1.0003
$P(A \cap B)$	0.6779	0.1253	0.5307	0.0003
$P(A \cap M)$	0.0000	0.0000	0.0000	0.0000
$P(B \cap M)$	0.0000	0.0000	0.0000	0.0000
Income Response Subsets	2000	2001	2002	2003
N (employed EAP)	21 455	21 189	20 508	20 156
(5) Exact Responses: P(A)	0.8737	0.7527	0.7055	0.6826
(6) Bounded Responses: P(B)	0.0951	0.1918	0.2284	0.2480
(7) Nonresponse: P(M)	0.0101	0.0519	0.0641	0.0683
(8) Complement: $(A \cup B \cup M)^{\mathsf{C}}$	0.0215	0.0036	0.0020	0.0011
Sum: (5) + (6) + (7) + (8)	1.0003	1.0000	1.0000	1.0000
$P(A \cap B)$	0.0003	0.0000	0.0000	0.0000
$P(A \cap M)$	0.0000	0.0000	0.0000	0.0000
$P(B \cap M)$	0.0000	0.0000	0.0000	0.0000

In the table, the column for 2000 is repeated for presentation purposes only, simply to show (1) how the transition from the OHS to the LFS proceeded, and (2) how all of the LFSs compare.

We can see from the table that the sum of the probabilities do not always add up to one; this is the first clue that something is amiss. The first form of error exists for the OHS97-LFS00, but only for the subset $P(A \cap B)$. That is, we sometimes jointly observe values for exact and bounded income for the same respondents in these public-use datasets, which should not be happening.

The findings for 1997 and 1999 are noteworthy because of the magnitude of the error in the data, at 68 and 53 percent, respectively (obtained from the "Sum" row in the table). For both years, these numbers match the percentage of actual income observations in the survey. This suggests that for each exact income observation, there is also a bounded observation. It is unclear why this is the case, or what motivation Statistics SA could possibly have had in doing this. One potential reason is that it is not a form of error at all, but rather that the survey organisation intentionally did this for some reason (it was not apparent from a reading of the survey organisation's accompanying literature and metadata whether or why this was done).

In order to investigate this further, we checked the consistency between the exact values that were also observed as brackets by transforming actual income into a new monthly income variable, and then converting that variable into a bracketed variable with the same bounds as the SSA's bounded variable. The result was that about 85 percent in 1997 and 99 percent of actual income observations in 1999 were in the correct monthly income bracket. For 1998, only 16 percent of actual income observations were in the correct bracket. While it is true that the extent of this error is mitigated to some extent when there is a match between the variables, the existence of two data points on income for the same person should never, as a rule, exist.

We do not observe this form of error for the other possible subsets, namely $P(A \cap M)$ or $P(B \cap M)$, in any of the datasets. This is unsurprising, for the actual placement of the "Don't Know" and "Refuse" options in the public-use dataset is as an option in the bounded income variable, making it impossible to confuse these subsets (when they enter the data electronically).

It is clear from the table, though, that SSA really improved their performance on this dimension of the problem over time, with this form of error dropping to zero by the LFSs. That said, the LFS2000–2003 all have non-zero complements to $P(A \cup B \cup M)$, which ought to no longer exist given that the income question had specific response options for don't know and refuse.

Consequently, a second form of error exists, and is non-zero in each LFS dataset. It is substantial in the OHS 1999 and LFS 2000, at approximately 2.5 and 2 percent, respectively, of the sample of employed economically active individuals.

The first type of error discussed for these datasets can easily be dealt with by generating a new derived income variable from the combined actual and interval variables in the raw data, and overwriting the bracketed responses with the exact responses. The rationale for doing this is that exact responses are preferred to bounded responses from an information content point of view (see Schwartz and Paulin, 2000). For the second type of error, we deal with it differently across the survey years: the observations are kept in the OHS 1999 because they are confounded with refusals; but they are omitted for imputation purposes from the LFS, where the nonrespondent subset is fully defined by don't know and refusals. However, we will evaluate and impute these response types separately in the analysis below to examine their distribution.

3.2 The Imputation Algorithm

There are several important steps required for the development of appropriate multiple imputation methods. These include:

- Correctly characterising the nature of the missing data, called the "missingness" mechanism. Little and Rubin (2002, 4-8) identify several such patterns, including univariate nonresponse, multivariate nonresponse (e.g. item nonresponse and unit nonresponse), monotone missing (e.g. attrition in longitudinal studies), general patterns of missing data (e.g. item nonresponse on many variables in a single dataset), file matching missing data problems, and latent-variable patterns with variables that are never observed. An important relationship exists between the pattern of missing data and the imputation procedure, with univariate and monotone missing data patterns allowing for the simplest imputation algorithms to be implemented (White, Wood and Royston, 2007).

- Based on the missing mechanism, choosing an appropriate multiple imputation algorithm. An important requirement of this choice is ensuring that the imputation method is "proper", which means that it must account for uncertainty in the parameters of the imputation model (White, Royston and Wood, 2011). This is necessary because Rubin's Rules for combining datasets only yield valid standard errors if the imputations adequately reflect the uncertainty in drawing values for the missing data.

- Specifying the imputation model: variable selection. As White, Royston and Wood (2011) point out, covariates for each prediction equation in the imputation algorithm have to be carefully chosen to help increase the plausibility of the missing (coarsened) at random assumption. Van Buuren, Boshuizen and Knook (1999) suggest that variable selection ought to include:
  - Variables that are required in the complete data model of interest;
  - Variables that appear to determine missingness;
  - Variables that explain a considerable amount of the variance of the target variable, which helps to reduce the uncertainty of the imputations.

- Specifying the imputation model: model form. An important concept in the imputation literature is the idea of a "congenial" imputation model. White, Royston and Wood (2011) state that instead of aiming to find the true imputation model, an alternative approach relies on finding an imputation model that is congenial to the analysis model but not necessarily correctly specified. In this way, inference on multiply imputed data can approximate maximum likelihood estimates (for large numbers of imputations) (ibid, 385).

- Choosing sufficiently large numbers of multiple imputations for the missing data in order to reflect the uncertainty present in the imputation process. Traditional multiple imputation theory used the oft-cited rule-of-thumb of five imputations, but more recent studies suggest that many more multiple imputations may be needed – in the order of one hundred for certain applications (Graham, Olchowski and Gilreath, 2007).

- Conducting complete-case analysis from multiply imputed data using the correct combination rules. Depending on the problem under investigation, these combination rules may differ to Rubin's Rules (Reiter and Raghunathan, 2007).

- Testing the sensitivity of the results. This can be done in different ways, since each step described above imposes a certain structure on the imputation process, the sensitivity of which can be investigated. Carpenter, Kenward, and White (2007) use a weighting approach after imputation to test the validity of the MAR assumption for each imputed dataset. However, this requires a specific model for how imputations depart from MAR. Sensitivity analysis can also be conducted using an uncongenial imputation model, which Kenward and Carpenter (2007) suggest. This involves specifying an imputation model that differs from the analysis model. We incorporate this suggestion into the analysis below.

It is important to note that in this paper we are concerned with multiply imputing for coarse income data only, which sets the pattern of coarseness as univariate. Consequently, we are not interested in multivariate coarsening or the effect of coarse data on the earnings covariate vector. An important consequence of this is that the multiple imputation algorithms simplify tremendously because the process of drawing plausible values from the conditional distribution of each variable with coarse data is restricted by design to one conditional distribution – income.

Practically, this means our task is to develop a univariate multiple imputation algorithm. This has two implications: (1) it is no longer necessary to characterise the coarse data mechanism in a multivariate sense (e.g. to establish whether it is monotonic or a general multivariate coarse data pattern); and (2) it is no longer necessary to use a sequential regression multiple imputation approach to the problem because there is only one variable with coarse data[1]. For this purpose we utilise the interval regression-based mul-

[1] The two most common sequential imputation algorithms are variants of Van Buuren, Boshuizen and Knook's (1999) multiple imputation by chained equations (MICE) algorithm, and Raghunathan, Lepkowski, Van Hoewyk and Solenberger's (2001) sequential regression multiple imputation (SRMI) algorithm. Royston's (2004, 2005, 2007, 2009) imputation by chained equations (ICE) algorithm is similar in principle to Van Buuren et al's (1999) procedure, while Statacorp (2011) developed a flexible multiple imputation package that can perform monotonic multiple imputation, fully conditional specification procedures (such as MICE, ICE and SRMI), and explicit Bayesian algorithms that allow the user to specify prior and posterior distributions, amongst others. The algorithm

tiple imputation procedure developed by Royston (2007) and modified by Statacorp (2011).

3.3 Estimation and Inference from Multiply Imputed Data

Multiple imputation was suggested as a potential solution to missing data problems by Rubin (1976), and the rules for inference from multiply imputed datasets came to be known as Rubin's Rules. These essentially state that analyses of multiply imputed datasets should be conducted based on standard complete-data techniques, but parameter estimates must be combined across datasets.

Formally, Rubin's Rules are presented as follows (we follow Royston's (2004) exposition): Let $\hat{\theta}_m, W_m, m = 1, ..., M$ be M complete-data estimates and their associated variances for an estimated parameter θ. The mean of θ is then calculated as:

$$\bar{\theta}_M = \frac{1}{M}\sum_{m=1}^{M}\hat{\theta}_m. \tag{3}$$

The variance of θ has both a within component and a between component. The within component of the variance is:

$$\bar{W}_M = \frac{1}{M}\sum_{m=1}^{M}W_m. \tag{4}$$

The between component of variance is:

$$B_M = \frac{1}{M-1}\sum_{m=1}^{M}(\hat{\theta}_m - \bar{\theta}_M)^2. \tag{5}$$

Combining the within and between-components then leads to the formula for total variance:

$$T_M = \bar{W}_M + \frac{M+1}{M}B_M, \tag{6}$$

The reference distribution for confidence intervals and significance tests is a t distribution,

in Statacorp (2011) also has the functionality to be restricted to the type of univariate multiple imputation procedure utilised here.

$$(\theta - \bar{\theta}_M)T_M^{-1/2} \sim t_\nu,$$

with degrees of freedom,

$$\nu = (M-1)\left(1 + \frac{1}{M+1}\frac{\bar{W}_M}{B_M}\right)^2.$$

In the analysis below, we obtain parameter estimates for the marginal distribution of post-multiply imputed income using these rules for a variety of different parameters.

4 Results: Univariate Multiple Imputations for Coarse Income

In this section we conduct univariate multiple imputation for coarse income data. Our objective is to draw plausible values for both the bracketed and missing subsets in each survey year. The multiple imputation algorithm employed for this purpose is based on an interval regression procedure developed by Statacorp in *Stata Release 12* (2011). The algorithm allows for imputed draws to be restricted to the income bracket lower and upper bounds, and it simultaneously allows for imputed draws for missing data to be unrestricted. The sensitivity of estimates and inferences to a range of different specifications of the prediction equations of the imputation algorithm is tested. Four models are developed for this purpose:

1. Model 1: multiply imputing five times with an intentionally mis-specified covariate vector that includes gender and language as the only predictors. The purpose of doing this is to create a baseline set of imputations that provide insight into how badly things can go wrong due to covariate mis-specification.

2. Model 2: multiply imputing five times with prediction equations using covariates that explain the response process only (see Table 6 on page 38 for the variables in this model, and Daniels (2012) for a motivation for the use of this model). The purpose of doing this is to create an "uncongenial" set of imputations, in the sense that the imputation model differs from the intended analysis model (Kenward and Carpenter, 2007).

3. Model 3: Multiply imputing five times for univariate income with Mincerian earnings function covariates only. These include age and experience (including their squares), other personal characteristics variables (including race and gender, but not language), hours worked, occupation, trade union membership, industry, and province. The purpose of this model is to create a set of imputations that would be "congenial" to analysing earnings, even though variables that explain the response process are largely absent.

4. Model 4: multiply imputing five times using both Mincerian earnings equation covariates and response propensity covariates. On a-priori grounds, this algorithm is treated as first-best because it conforms to the recommendations of Van Buuren et al (1999, see section 3.2 for discussion).

4.1 Quantiles and Moments Across Four Imputation Models

The results for weighted univariate income parameter estimates for each imputation model are presented in Table 3. The table shows parameter estimates of the multiply imputed nominal employment income variables ("Yimp"), for each of the four imputation models discussed above and the estimation sample size ("Est.N") in each survey year. Quantile estimates are calculated post-imputation for each of m imputed income variables using Rubin's Rules (see equation [3] above). For this section, the variance of the estimates are omitted, but they will be evaluated in detail below in section 4.6[2].

[2] Note that the variance of a quantile has to be computed manually after m multiple imputations using Rubin's Rules (see equations [4] to [6] above). The total variance of a quantile contains only a between-imputation component of variance (see equation [5] above), but Rubin's total variance formula in equation [6] still has to be used to calculate the variance of a quantile because of the $(m + 1)/m$ adjustment for finite m.

Table 3: Quantiles of Four Different Models for Imputed Income

Year	Variable	min	p5	p10	p25	p50	mean	p75	p90	p95	p99	max	Est.N
1997	Yimp-model1	0	211	350	863	1 796	4 054	4 000	8 705	14 512	37 724	307 832	23 868
	Yimp-model2	0	204	350	804	1 700	3 688	3 665	7 871	12 918	33 526	202 582	23 303
	Yimp-model3	0	206	350	803	1 709	3 433	3 660	7 548	12 028	27 451	177 681	23 206
	Yimp-model4	0	201	348	800	1 656	3 287	3 516	7 278	11 457	26 572	127 069	22 805
1998	Yimp-model1	0	206	304	800	1 951	5 600	4 980	12 213	21 397	61 051	511 400	12 985
	Yimp-model2	0	201	300	772	1 809	5 210	4 673	11 532	19 980	54 850	598 968	12 574
	Yimp-model3	0	200	300	681	1 608	3 910	3 971	8 836	14 488	35 832	370 000	11 619
	Yimp-model4	0	200	300	652	1 586	3 756	3 803	8 270	13 741	33 601	370 000	11 356
1999	Yimp-model1	0	216	337	785	2 000	7 549	5 869	15 441	28 147	88 298	1 559 224	21 915
	Yimp-model2	0	213	311	700	1 757	6 376	4 970	13 008	23 760	72 413	1 522 138	20 365
	Yimp-model3	0	216	312	700	1 796	6 041	5 014	12 879	22 483	61 965	1 522 138	20 575
	Yimp-model4	0	200	300	678	1 702	5 697	4 738	12 137	21 297	56 636	1 522 138	19 562
2000	Yimp-model1	0	217	318	665	1 521	5 890	3 500	7 037	11 146	27 474	4 726 242	20 993
	Yimp-model2	0	217	304	652	1 500	5 824	3 486	6 965	10 934	25 572	4 726 242	20 734
	Yimp-model3	0	217	305	652	1 500	5 804	3 500	7 000	10 921	24 686	4 726 242	20 725
	Yimp-model4	0	216	300	652	1 500	5 678	3 358	6 611	10 157	23 446	4 726 242	20 538
2001	Yimp-model1	0	250	350	748	1 800	4 120	4 383	8 999	14 894	37 827	500 000	21 112
	Yimp-model2	0	248	350	700	1 738	3 751	4 000	8 161	13 277	32 644	500 000	20 486
	Yimp-model3	0	250	350	702	1 738	3 681	4 000	8 098	12 934	30 953	500 000	20 599
	Yimp-model4	0	242	350	700	1 700	3 471	4 000	7 855	11 972	28 095	500 000	20 156
2002	Yimp-model1	0	250	350	763	1 919	5 399	5 012	11 827	20 190	55 296	500 797	20 467
	Yimp-model2	0	250	350	737	1 800	4 896	4 957	11 010	19 021	45 871	396 532	19 834
	Yimp-model3	0	250	350	750	1 842	4 448	4 844	10 159	16 738	38 143	380 000	19 994
	Yimp-model4	0	250	350	701	1 800	4 122	4 580	9 618	15 558	34 388	380 000	19 549
2003	Yimp-model1	0	300	480	856	2 000	5 925	5 653	13 120	22 415	59 026	726 726	20 130
	Yimp-model2	0	300	477	846	2 000	5 300	5 145	12 161	20 446	51 422	321 882	19 599
	Yimp-model3	0	300	495	854	2 000	5 048	5 226	11 904	19 330	45 200	240 975	19 805
	Yimp-model4	0	300	472	818	2 000	4 697	5 000	11 027	17 980	40 299	212 935	19 359

Results from the table are discussed thematically. The following issues are of relevance:

- The difference in parameter estimates across imputation methods.
- The difference in the estimation sample size across imputation methods.
- The difference in the upper and lower tails of each distribution.

Evident from table 3 is that up until the median, the differences between the imputations are relatively trivial. This is expected, for we know that the probability of a bounded responses increases as income increases, so any difference in imputed draws for this subset will only make its presence felt higher up the income distribution. That said, an important feature of the imputation algorithm is that it limits the range of imputed draws to the bounds of each income category. For the highest income category, however, this is an open ended interval with no upper bound. Therefore, imputations for respondents in this group have no upper limit.

At the top of the income distribution, we see substantial differences between the distributions. At the 99th percentile, the OHS 1999 has the widest range between the four imputation models. The mis-specified method of model 1 leads to substantially higher estimates than any other model. The differences between distributions in model 2 (that has response propensity covariates) and model 3 (that has earnings function covariates) is also substantial, but the difference in estimates between model 3 and the first-best imputation model 4 (which combines response propensity and earnings function covariates) is much lower.

In fact, in every survey year and for every quantile other than the minimum, the first-best imputation model always generates distributions with the lowest estimates. The importance of this is particularly stark for the maximum values in each distribution. Important to note here is that in survey years where an exact income value is extreme, such as in 1999 and 2000, the imputed values rarely exceed this outlier, except for the mis-specified imputation model one in 1999, where an imputed draw is larger than the maximum in that year. But there is nothing generalisable from this observation, for in 2001 where an exact income value also represents the maximum, the imputation model one does not exceed it. The relationship between outliers in the observed distribution and multiple imputation is therefore important to be aware of.

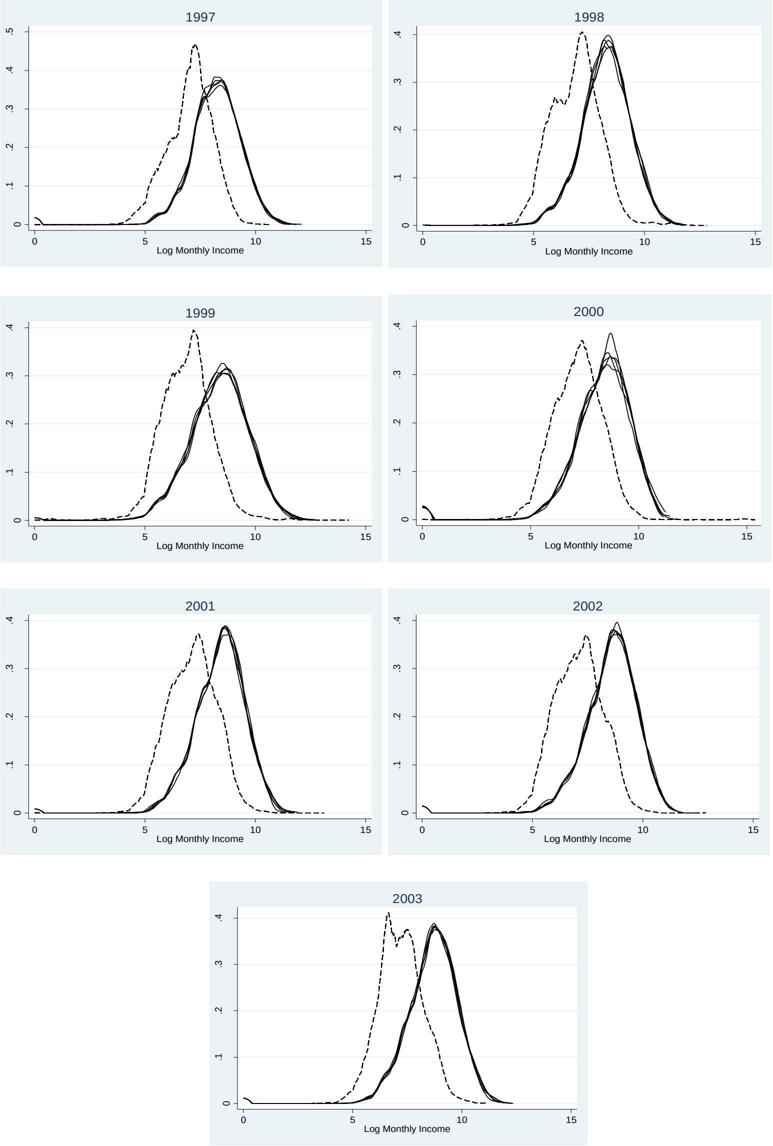
The differences between the four imputation models at the maximum are substantial in 1997, 1998, 2002, and 2003. This suggests that specification of the imputation algorithm is most significant to the upper tail of the income distribution. The fact that the model 4 estimates are the lowest for each parameter across the entire distribution suggests that covariate selection based on explaining both the outcome variable of interest (income) and the response process leading to coarse data (response propensities), is crucial for plausible draws of income, but even more important the highest income earners.

However, it is not clear that a congenial imputation model that only focuses on earnings covariates (model three) is substantially worse than model four. Model two is slightly more volatile across the survey years, suggesting that choosing covariates that explain the response process alone is not an optimal way of specifying multiple imputation algorithms. Finally, the reduction in the estimation sample size for model 4, although relatively modest, is nevertheless an important limitation associated with increasing the number of covariates in the prediction equations.

4.2 The Distribution of Multiply Imputed Bounded Income Values

In this section we compare the subsets of multiply imputed income. We restrict the analysis initially to the first-best imputation model only. The kernel densities of the five multiply imputed bounded income distributions are presented in Figure 1. The density for exact income responses is on the same graph. The solid lines represent the bounded distributions and the dashed line the continuous distribution for exact responses.

Figure 1: Multiply Imputed Bracketed Income Compared to Observed Continuous Income: 1997-2003



1997

1998

1999

2000

2001

2002

2003

We can see from figure 1 that the densities of imputed draws for the bracketed subset are always to the right of the actual income response distribution. This is entirely expected from the analysis in Daniels (2012), where it was shown that the probability of a bounded income response increases as income increases.

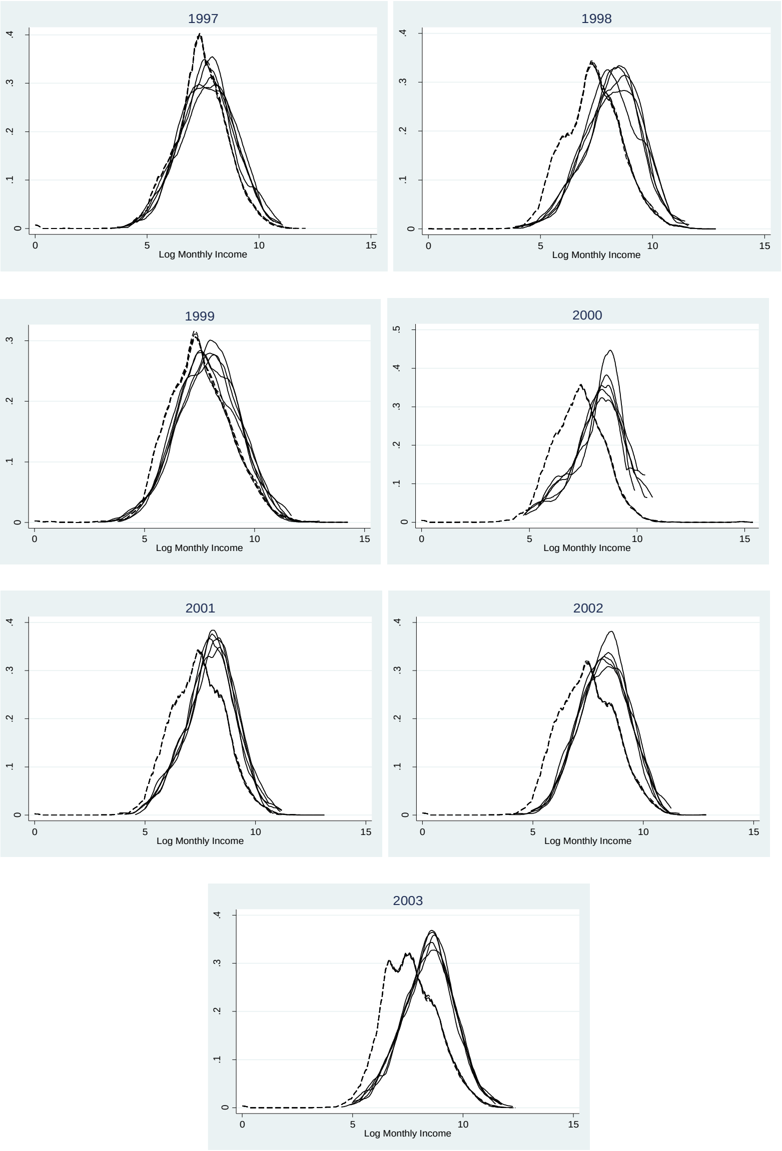
The densities for each of the five imputed draws are very similar, and generally have similar skewness and kurtosis. This is to be expected given the bounds of the brackets, which restrict where in the distribution the draws can be made. An important observation concerns the maxima of the imputed draws for the bracketed subset of income respondents. In 1997 and 2003 we see clearly that the maximum monthly income value in the data is generated by the imputed draws for bounded income.

It is also apparent that the minimum income values are determined by respondents who answer the bracketed section of each questionnaire. It should be remembered that the lowest bracket in each questionnaire is zero. And in each survey year we observe a non-zero count of such responses. This is highest in 2000, but is also noticeable in 1997, 2001-2003, where it clearly affects the kernel densities. The existence of zero values for employee income is not unreasonable given the fact that the income question asks respondents about their labour market activities in the week preceding the interview, during which respondents could be earning no income.

4.3 The Distribution of Multiply Imputed Missing Income Values

The kernel densities of multiply imputed draws for the nonresponse subset (combining unspecifieds, don't know and refusals as appropriate to the survey year) of observations are compared to the observed responses (bounded and continuous) in figure 2. As before, each of the five multiply imputed income distributions are plotted on the same graph for each year. The densities for imputed draws of missing income observations are the solid lines while observed income has dashed lines.

Figure 2: Multiply Imputed Missing Income Compared to Observed (Multiply Imputed Bracket & Continuous) Income: 1997-2003



We can see from this figure that the distribution of imputed missing values changes over time, relative to the distribution of observed responses. In 1997 the densities for the missing income respondents generally overlaps that of the observed respondents. This suggests that respondents who didn't answer the income question had similar predicted values of income compared to respondents who did provide an answer to the question, based on observables in the public-use dataset. That begins to change immediately after 1997, however, where in 1998 it becomes clear that the missing subset of respondents had predicted income values discernibly more to the right than the observed subsets of income respondents.

The location of the densities for the missing subset of observations gradually moves further to the right over time. To explain this trend, it is noteworthy to remember that we are observing the *nominal* distribution of monthly income over time. It is therefore reasonable to expect that the distribution of income in the population itself would shift to the right over the time frame.

4.4 The Distribution of Multiply Imputed Refusals and Don't Know Income Values

In this section we evaluate the distributions of multiply imputed refusals and don't know income values. The time frame is restricted to 2000 and beyond, since these response options only appear in the questionnaires from 2000 onwards. The kernel densities for the multiply imputed draws of refusals are plotted with a solid line while draws for don't know responses are plotted with dashed lines. Because imputed draws for refusals and don't know responses are of particular interest, we compare the four multiple imputation models against each other. In figure 3, the mis-specified imputation method (model 1) is on the left hand side while the first-best imputation method (model 4) is on the right hand side.

It is evident from figure 3 that there is now a lot more variation between the imputed draws for each response group, and there are very different inferences about the distribution of don't know and refuse responses depending on which multiple imputation method is used. According to model one, the two groups are nearly indistinguishable, whereas in model four they are always very different. The densities of imputed income draws for refusals always lie to the right of the don't know responses. This shows a clear advantage of correctly specifying multiple imputation algorithms.

Figure 3: Multiply Imputed Missing Income: Refusals Compared to Don't Know: 2000-2003

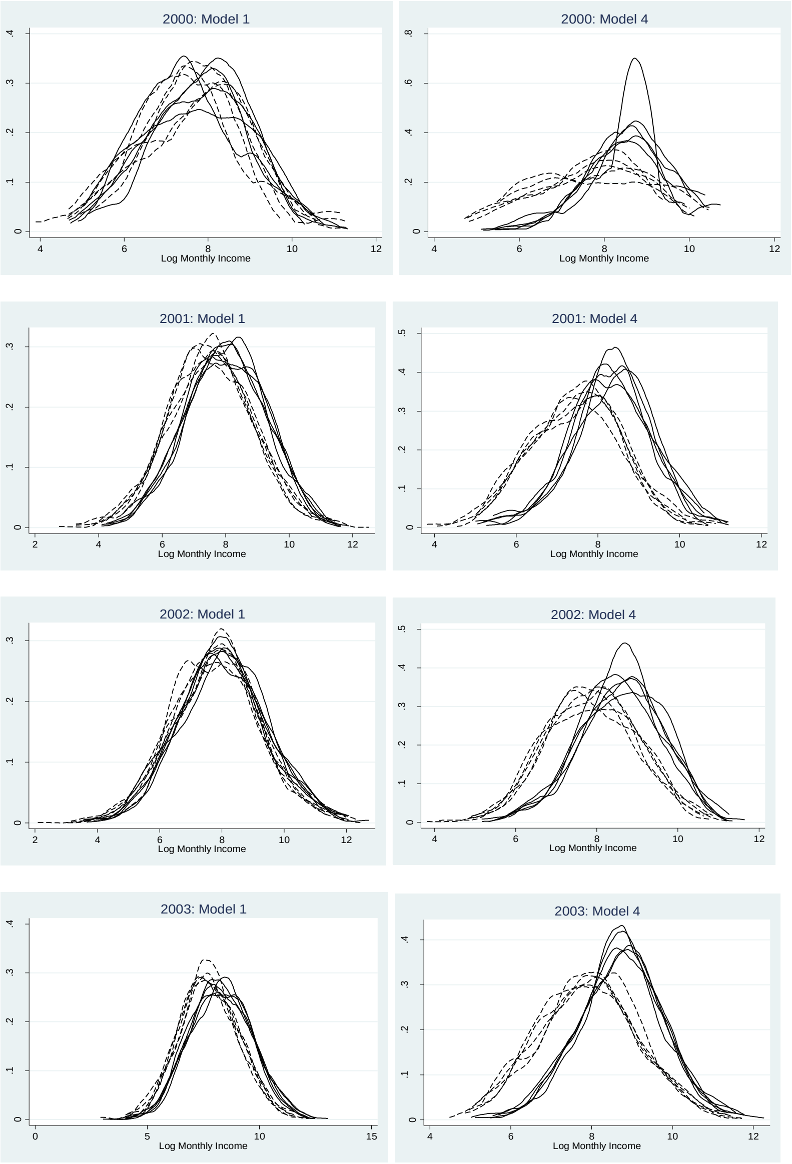
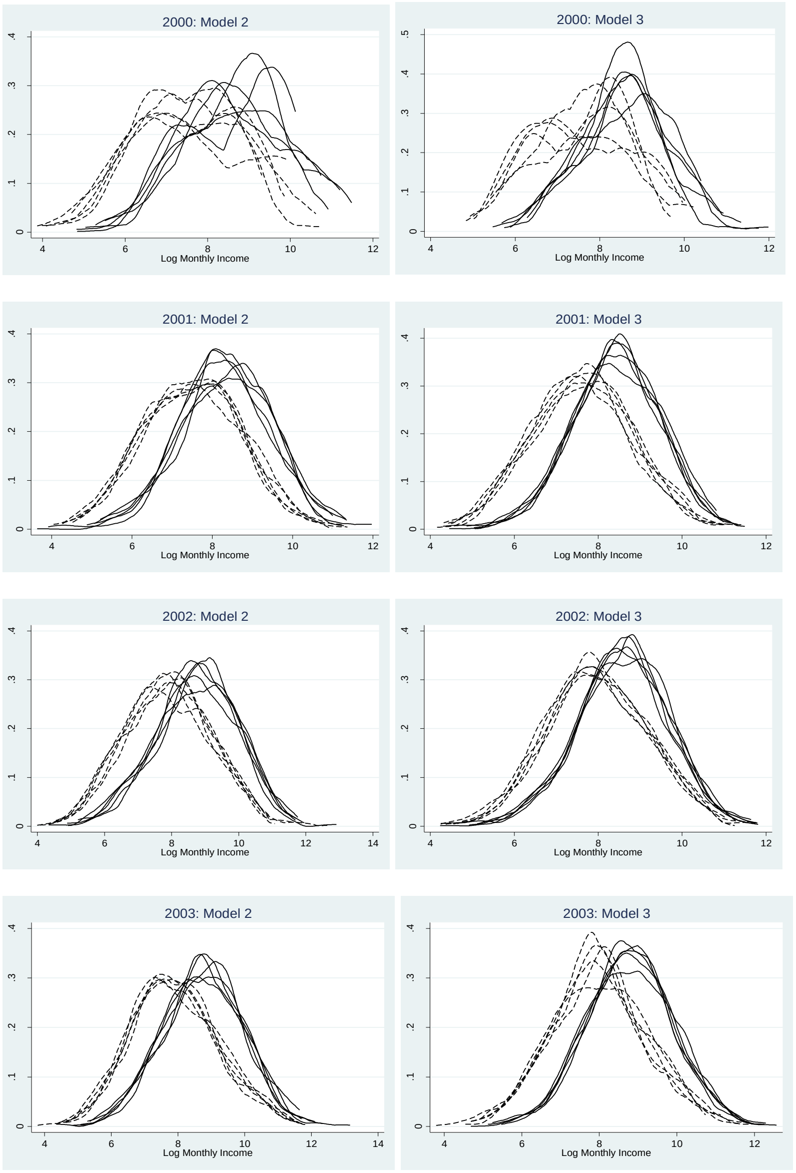


Figure 4: Refusals Compared to Don't Know: Response Propensity (Model 2) and Earnings Function (Model 3) Imputations: 2000-2003


2000: Model 2

2000: Model 3

2001: Model 2

2001: Model 3

2002: Model 2

2002: Model 3

2003: Model 2

2003: Model 3

To evaluate the sensitivity of this finding, we now compare the results for multiple imputation models 2 and 3 against each other. Figure 4 presents the densities where refuse responses are the solid lines while don't know responses are the dashed lines.

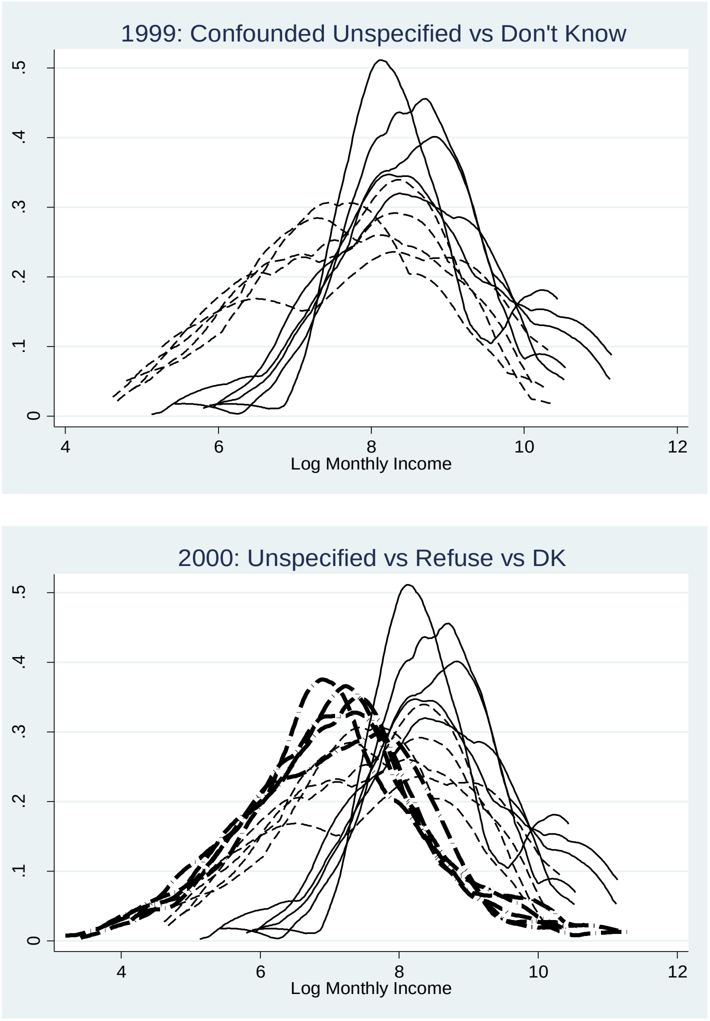
We can see from figure 4 that regardless of whether the multiple imputation algorithm is specified with response propensity covariates only, or whether it is specified with earnings function covariates, the imputed draws for don't know and refuse subsets of the income distribution show very different distributions. The fact that both models predict this difference is unsurprising because some of the response propensity covariates were chosen precisely because they're correlated with income.

4.5 Unspecified Responses as a Source of Error

In this section we isolate two survey years where unspecified responses represent a significant source of error, namely 1999 and 2000. Unspecified responses in 1999 are confounded with refusals; they consequently enter into the multiple imputations models discussed above. However, in 2000 unspecified responses represent a source of error only because don't know and refuse responses complete the nonresponse possibilities. Therefore, these responses are not imputed in models 1 through 4 above. However, in this section we conduct a new multiple imputation exercise for the LFS 2000 that is identical to model 4 above, but that does multiply impute values for unspecified responses. We then evaluate the densities of these unspecified responses compared to the other nonresponse subsets.

Table 1 on page 10 presents the subsample sizes for unspecified responses. We now want to compare the multiply imputed draws for these responses against the imputed draws for don't know responses in 1999, and against both don't knows and refusals in 2000. Figure 5 presents the results. In 1999, the densities for unspecified income draws are the dashed lines, while the solid lines represent don't know responses. In 2000, the densities for unspecified income draws are the bold dashed lines, whereas refusals are the solid lines and don't know the narrower dashed lines.

Figure 5: Unspecified Response Error Imputations: 1999 and 2000



From figure 5 it is clear that unspecified responses are substantially different to identified nonresponse groups in both 1999 and 2000. In 1999, if the unspecified responses were only refusals, then we would expect the distribution of these responses to lie to the right of the imputed don't know densities, as they do for every survey year in figures 3 and 4. However, they are much more widely spread across the income distribution than refusals.

The same is true in 2000, when there is no longer confounding with refusals. Here, the densities for the imputed unspecified responses are spread across a much larger range than either the don't know or refuse imputations. This suggests that processing error is a completely different error mechanism to nonresponse on the income question, and should consequently not enter multiple imputation algorithms that do not explicitly account for the very different properties of this component of error.

4.6 Stability of Parameter Estimates as the Number of Multiple Imputations Increase

The final section of this paper evaluates the stability of parameter estimates of imputed income as the number of imputations increase from two to five to twenty. We conduct multiple imputations using the specification of model 4 only. A-priori, we know that there is not much variation in imputed draws below the median of monthly income from previous analysis (see Table 3 on page 20). However, above this level there is more scope for variation. In particular, the largest (open-ended) income bracket as well as the distribution for imputed refusals and don't know responses should be considered to be highly variable given the analysis above. We therefore need to establish the bounds of sensitivity due to the number of multiple imputations conducted. Tables 4 and 5 present the results of this exercise.

Table 4: Quantile Estimates of Imputed Income as Number of Imputations Increase

Yr & # Imps	p10	p25	p50	mean	p75	p90	p95	p99	max	wgt.sum	Est.N
97 m=2	348	800	1 652	3 291	3 550	7 285	11 487	26 409	124 035	25 097 196 736	22 805
97 m=5	348	800	1 656	3 287	3 516	7 278	11 457	26 572	127 069	25 067 172 581	22 805
97 m=20	348	800	1 661	3 310	3 523	7 271	11 493	26 856	157 552	25 241 903 829	22 805
98 m=2	300	652	1 586	3 766	3 834	8 394	13 808	32 589	370 000	22 605 113 545	11 356
98 m=5	300	652	1 586	3 756	3 803	8 270	13 741	33 601	370 000	22 547 061 243	11 356
98 m=20	300	652	1 592	3 809	3 826	8 435	13 990	34 417	370 000	22 864 444 500	11 356
99 m=2	300	674	1 704	5 651	4 712	11 998	20 982	57 871	1 522 138	43 505 765 737	19 562
99 m=5	300	678	1 702	5 697	4 738	12 137	21 297	56 636	1 522 138	43 867 855 872	19 562
99 m=20	300	674	1 702	5 650	4 703	12 084	21 026	55 297	1 522 138	43 499 371 526	19 562
00 m=2	300	652	1 500	5 683	3 350	6 654	10 076	22 779	4 726 242	50 081 776 261	20 538
00 m=5	300	652	1 500	5 678	3 358	6 611	10 157	23 446	4 726 242	50 044 395 951	20 538
00 m=20	300	652	1 500	5 686	3 349	6 635	10 103	23 158	4 726 242	50 112 869 667	20 538
01 m=2	350	700	1 700	3 481	4 000	7 936	12 086	27 635	500 000	28 759 747 602	20 156
01 m=5	350	700	1 700	3 471	4 000	7 855	11 972	28 095	500 000	28 683 413 421	20 156
01 m=20	350	700	1 704	3 489	4 000	7 951	12 019	28 640	500 000	28 826 536 243	20 156
02 m=2	350	700	1 800	4 161	4 591	9 837	15 897	34 629	380 000	35 123 620 901	19 549
02 m=5	350	701	1 800	4 122	4 580	9 618	15 558	34 388	380 000	34 800 753 362	19 549
02 m=20	350	704	1 800	4 153	4 582	9 662	15 494	34 306	380 000	35 060 187 137	19 549
03 m=2	471	828	2 000	4 685	5 000	11 119	18 175	39 574	145 035	42 606 474 187	19 359
03 m=5	472	818	2 000	4 697	5 000	11 027	17 980	40 299	212 935	42 717 106 246	19 359
03 m=20	470	813	2 000	4 732	5 001	11 215	18 300	40 466	225 885	43 033 850 802	19 359

Parameter estimates in Table 4 are calculated as the mean of the two, five and twenty multiply imputed monthly income variables in the each respective datasets, as per equation 3 of Rubin's Rules. Evident from the table is that quantile estimates are almost identical below the median. For the mean of monthly income, the estimates are also very close across the two, five and twenty imputations for each survey year. In fact, this observation holds for every quantile including the maximum in every survey year. Even when we sum up all of the observations for monthly income to create a population-based estimate of the total monthly income earned by employees in South Africa, we can see that estimates do not differ substantially.

The coefficient of variation of these estimates is presented in table 5. Given that the means of parameter estimates are stable over two, five and twenty imputations–as presented in table 4–the coefficient of variation is informative about the magnitude of the inflation in the variance observed as the number of imputations increase.

We can see from the table that the ratio of the standard deviation to the mean is very small across every quantile and moment as the number of imputations increase. The largest values for the coefficient of variation are all found in the maximum column, for the survey years 1997 and 2003. Even here though, the numbers are less than 0.5. Aside from these larger values, every other estimate of the coefficient of variation is always below 0.1.

Despite the small magnitude of these coefficients, an important observation is the fact that they do not simply reduce in size as the number of imputations increase. This prevents any strong conclusions about the relationship between the number of imputations and its impact on inference. Two contributing factors to this finding are that (1) the percentage of missing observations is small (at between 3-5 percent for each survey year), and (2) the range of the bounded subset of observations is restricted through the imputation algorithm to lie within the lower and upper bound of each income bracket, thereby formulaically reducing the variance for imputed draws for all but the highest, open-ended income bracket.

Table 5: Coefficient of Variation of Quantiles and Moments as Number of Imputations Increase

Yr & # Imputations	p10	p25	p50	mean	p75	p90	p95	p99	max	sum	N
97 m=2	0.0000	0.0000	0.0107	0.0054	0.0062	0.0029	0.0047	0.0100	0.0344	0.0052	22805
97 m=5	0.0026	0.0000	0.0063	0.0137	0.0045	0.0159	0.0229	0.0516	0.2556	0.0137	22805
97 m=20	0.0013	0.0000	0.0077	0.0116	0.0076	0.0129	0.0171	0.0382	0.3134	0.0116	22805
98 m=2	0.0000	0.0000	0.0125	0.0272	0.0245	0.0393	0.0553	0.0000	0.0000	0.0272	11356
98 m=5	0.0000	0.0000	0.0090	0.0104	0.0101	0.0160	0.0295	0.0280	0.0000	0.0104	11356
98 m=20	0.0000	0.0000	0.0072	0.0228	0.0136	0.0275	0.0401	0.0615	0.0000	0.0228	11356
99 m=2	0.0000	0.0000	0.0033	0.0229	0.0036	0.0071	0.0204	0.0585	0.0000	0.0229	19562
99 m=5	0.0000	0.0073	0.0136	0.0229	0.0096	0.0180	0.0263	0.0744	0.0000	0.0229	19562
99 m=20	0.0000	0.0145	0.0044	0.0179	0.0131	0.0157	0.0242	0.0469	0.0000	0.0179	19562
00 m=2	0.0000	0.0000	0.0000	0.0061	0.0211	0.0114	0.0107	0.0484	0.0000	0.0062	20538
00 m=5	0.0000	0.0000	0.0000	0.0068	0.0082	0.0075	0.0205	0.0480	0.0000	0.0068	20538
00 m=20	0.0000	0.0000	0.0000	0.0059	0.0138	0.0099	0.0185	0.0357	0.0000	0.0059	20538
01 m=2	0.0000	0.0000	0.0000	0.0087	0.0000	0.0101	0.0100	0.0547	0.0000	0.0087	20156
01 m=5	0.0000	0.0000	0.0000	0.0123	0.0000	0.0138	0.0041	0.0558	0.0000	0.0122	20156
01 m=20	0.0000	0.0000	0.0044	0.0091	0.0000	0.0087	0.0099	0.0344	0.0000	0.0091	20156
02 m=2	0.0000	0.0000	0.0000	0.0046	0.0029	0.0094	0.0179	0.0152	0.0000	0.0046	19549
02 m=5	0.0000	0.0025	0.0000	0.0095	0.0090	0.0140	0.0114	0.0325	0.0000	0.0095	19549
02 m=20	0.0000	0.0070	0.0000	0.0107	0.0121	0.0127	0.0154	0.0293	0.0000	0.0107	19549
03 m=2	0.0286	0.0214	0.0000	0.0006	0.0000	0.0067	0.0055	0.0081	0.0360	0.0006	19359
03 m=5	0.0185	0.0037	0.0000	0.0125	0.0000	0.0103	0.0016	0.0540	0.0652	0.0125	19359
03 m=20	0.0162	0.0099	0.0000	0.0113	0.0005	0.0165	0.0167	0.0376	0.3968	0.0113	19359

For the highest, open-ended income bracket, we saw that specification of the prediction equation in the imputation algorithm is important for reducing the right skewness of the upper tail. Since parameter estimates in tables 4 and 5 use both response propensity and earnings function covariates in the the model, the variance even in the upper tail of the distribution is relatively low.

The overall conclusion from this analysis is that stability in the point estimates of parameters of multiply imputed income is achieved with as little as two multiple imputations.

5 Conclusion

In this paper we conducted univariate multiple imputation for coarse subsets of the employee income distribution in South African household surveys from 1997-2003. During this time, the employee income question itself evolved, shedding greater light on the coarse response mechanism. The coarse data framework was very useful in guiding the approach not only to the imputation algorithm, where an important implication was restricting the range of the imputed draws to lie within each income bracket, but also to the treatment of unspecified responses when they were identified as a source of survey error. This is one of the major advantages of the coarse data framework: it encourages an explicit approach to the characterisation of the response mechanism, which then leads to clear rules about what can and cannot be accommodated in the imputation step.

For processing error, the fact that two variables are released in the public-use dataset – one for actual income responses and one for bracketed responses – implies that there is a non-zero chance of error between these variables that needs to be addressed when it exists. We identified two types of survey errors: one where duplicate income responses were identified for the same individual, and another where unspecified responses were present in the data even when response options that complete the missing data subset were present in the questionnaire (i.e. don't know and refusals). The solution to the first type of error was to create a new variable for income that overwrites the duplicate records of bounded income with the actual income values. However, for the second type of error, there was no simple solution because the problem ought not to exist for the subsample of interest (employed economically active individuals). Hence these observations were not imputed in the main analysis and analysed separately instead.

An important relationship that repeatedly presented itself in each section of this paper was that of the relationship between questionnaire design and the resulting data structure. This made the analytical task iterative to an extent more than complex, for it required careful data checks and question wording and sequencing checks that mandated a fastidious and detail-oriented approach to the problems and interpretation of the results. An overall lesson learnt from this analysis is that it is incumbent upon researchers to be absolutely meticulous in their data preparation, imputation, estimation and analysis tasks when working with micro datasets.

The univariate approach to multiple imputation utilised here allowed for very specific sensitivity analyses to be performed. Four different specifications of the imputation models provided the basis for sensitivity analysis to mis-specification in the imputation algorithm. We used four different models for this purpose: a mis-specified algorithm (model 1), one that explained the response process only (model 2), one that explained income itself (model 3), and a final one that combined covariates from model 2 and 3. It was this fourth model that was chosen as the first-best model, given the recommendations for covariate selection of Van Buuren et al (1999). The main limitation with this model was a reduction in the estimation sample size due to the greater prevalence of covariate missing data compared to the other models.

The advantage of incorporating predictors for the response process in the imputation algorithm as well as earnings covariates was that it evidently reduced the right-skewness of the imputed monthly income values. The plausibility of imputed draws for the highest, open-ended income bracket, the refusals, don't know and unspecified response groups, was clearly affected by covariate selection in the imputation process. The fact that the first-best model reduced these values relative to the other three specifications suggests there is considerable merit to paying close attention to the response process in multiple imputation algorithms and not simply to predictors of the outcome variable.

This has important implications for more sophisticated multiple imputation exercises that seek to impute for covariate coarse data too, for it suggests that each variable with coarse observations needs: (1) a model of the coarse data mechanism for that variable (this would include checks for additional forms of survey error); (2) an analysis of the factors explaining the response process for that variable; and (3) appropriate prediction equations for that variable, which include covariates that explain both the response process and the outcome variable of interest.

Appendix: Response Propensity Model Predictors

See next page.

Table 6: Response Propensity Model Predictors

Variable	Rationale for inclusion	Attribute of respondent being tested
Household head	If respondent is HHH, more likely to know about incomes in the hh	Cognitive Burden (CB)
Self reporter	If a respondent is SR, more likely to know exact income	CB
Cohabiting status	If respondent in a cohabiting relationship, more likely to know spouse or partner's income	CB
HH composition	Tests effects of number of kids ($\leq$15) & adults (16-64) relative to the # of seniors ($\geq$65; reference group) in hh. The expected sign here is that an additional adult should increase CB of reporting	CB
Household size	The larger the size of hh, the less likely respondent knows all incomes	CB
Male	Personal characteristics of respondent or proxy	Personal Characteristics (PC)
Race	Personal characteristics of respondent or proxy	PC / CI / WD
Education	Education category of respondent or proxy	PC / CI
First Language (1)	Dummies for 11 official languages in SA. Captures possible socio-cultural influence to disclose income, though effects ambiguous	Willingness to disclose (WD)
First Language (2)	Simplified from above to four main SA first languages: Zulu, Xhosa, Afrikaans & English. All others combined into "Other"	WD
Wealth approximation	Derived from interaction of home ownership dummy with dwelling type: (1) Owned formal dwelling, including brick house, semi-detached house, flat or retirement unit (2) Unowned formal dwelling, same dwelling types as above (3) Sub-let room or dwelling, including room in main dwelling or structure in backyard (shack or room), not interacted with ownership (4) Mud hut or shack in squatter settlement, not interacted with ownership	Correlate of Income (CI)
Expenditure	Total household expenditure: continuous in 97,98 & 00; categorical in 99, 2001-2003	CI
Owns vehicle	Dummy for whether respondent owns vehicle or not. Reflects stock of wealth	CI
Felt unsafe in neighbourhood	If respondent feels unsafe, less likely to disclose income (only available in 97 & 98)	WD
Urban	Testing the effect of location. Has possible effect on willingness to disclose income	WD

References

[1] Allison, P.D., 2000, “Multiple imputation for missing data: A cautionary tale”, *Sociological Methods & Research*, 28, 301-309

[2] Ardington, C., Lam, D., Leibbrandt, M., and Welch, M., 2006, “The sensitivity of estimates of changes in post-Apartheid poverty and inequality to key data imputations”, *Economic Modelling*, 23, 822-835

[3] Carpenter, J.R., Kenward, M.G., White, I.R., 2007, “Sensitivity analysis after multiple imputation under missing at random: a weighting approach”, *Statistical Methods in Medical Research*, 16: 259-275

[4] Daniels, R.C., 2012, “Questionnaire design and response propensities for employee income micro data”, Southern Africa Labour & Development Research Unit (SALDRU) Working Paper Number 89, Cape Town: SALDRU

[5] Daniels, R.C., 2008, “The income distribution with coarse data”, Cape Town: Economic Research Southern Africa Working Paper Number 82

[6] Ghosh-Dastidar, B. and Schafer, J.L., 2003, “Multiple edit / multiple imputation for multivariate continuous data”, *Journal of the American Statistical Association*, 98(464): 807-817

[7] Graham, J.W., Olchowski, A.E., Gilreath, T.D., 2007, “How many imputations are really needed? Some practical clarifications of multiple imputation theory”, *Preventative Science*, 8: 206-213

[8] Heeringa, S.G., 1995, “Application of generalized iterative Bayesian simulation methods to estimation and inference for coarsened household income and asset data”, *The Proceedings of the Section on Survey Methods*, American Statistical Association, 42-51

[9] Heeringa, S.G., Little, R.J.A. and Raghunathan, T.E., 2002, “Multivariate imputation of coasened survey data on household wealth”, in Groves, R.M., Dillman, D.A., Eltinge, J.L., Little, R.J.A. (eds), *Survey nonresponse*, New Jersey: Wiley

[10] Heitjan, D.F., 1989, “Inference from grouped continuous data: A review”, *Statistical Science*, 4(2), 164-179

[11] Heitjan, D.F., 1994, “Ignorability in general incomplete data models”, *Biometrika*, 81, 701-708

[12] Heitjan, D.F. and Basu, S., 1996, “Distinguishing “Missing at Random” and “Missing Completely at Random””, *The American Statistician*, 50(3), 207-213

[13] Heitjan, D.F. and Rubin, D.B., 1990, "Inference from coarse data via multiple imputation with application to age heaping", *Journal of the American Statistical Association*, 85(410), 304-314

[14] Heitjan, D.F. and Rubin, D.B., 1991, "Ignorability and coarse data", *The Annals of Statistics*, 19(4), 2244-2253

[15] Kenward, M.G. and Carpenter, J., 2007, "Multiple imputation: Current perspectives", *Statistical Methods in Medical Research*, 16: 199-218

[16] Little R.J.A. and Rubin, D.B., 2002, *"Statistical analysis with missing data, Second edition"*, New Jersey: John Wiley and Sons

[17] Raghunathan, T.E., Lepkowski, J.M., Van Hoewyk, J. and Solenberger, P., 2001, "A multivariate technique for multiply imputing missing values using a sequence of regression models", *Survey Methodology*, 27(1), 85-95

[18] Reiter, J.P. and Raghunathan, T.E., 2007, "The multiple adaptations of multiple imputation", *Journal of the American Statistical Association*, 102(480): 1462-1471

[19] Royston, P., 2004, "Multiple imputation of missing values", *The Stata Journal*, 4(3), 227-241

[20] Royston, P., 2005, "Multiple imputation of missing values: Update", *The Stata Journal*, 5(2), 188-201

[21] Royston, P., 2007, "Multiple imputation of missing values: further update of ice, with an emphasis on interval censoring", Stata Journal 7(4), 445-464

[22] Royston, P., 2009, "Multiple imputation of missing values: Further update of ice, with an emphasis on categorical variables", Stata Journal 9(3), 466-477

[23] Rubin, D.B., 1976, "Inference and missing data", *Biometrica*, 63, 581-592

[24] Rubin, D.B., 1987, *"Multiple imputation for nonresponse in surveys"*, New York: Wiley

[25] Rubin, D.B., 1996, "Multiple imputation after 18+ years", *Journal of the American Statistical Association"*, 91(434), 473-489

[26] Schafer, J.L., 1999, "Multiple imputation: A primer", *Statistical Methods in Medical Research*, 8: 3-15

[27] Schwartz, L. and Paulin, G., 2000, "Improving Response Rates to Income Questions", *American Statistical Association (ASA) Section on Survey Research Methods*, Proceedings, 965-970

[28] StataCorp, 2011, *Stata Multiple Imputation Reference Manual: Release 12*, College Station: StataCorp LP

[29] Van Buuren, S., Boshuizen, H.C., and Knook, D.L., 1999, "Multiple Imputation of Missing Blood Pressure Covariates in Survival Models", *Statistics in Medicine*, 18, 681-694

[30] Vermaak, C., 2010, "The impact of multiple imputation of coarsened data on estimates of the working poor in South Africa", World Institute for Development Economics Research (WIDER) Working Paper No. 2010/86, Helsinki: WIDER

[31] White, I.R., Daniel, R. and Royston, P., 2010, "Avoiding bias due to perfect prediction in multiple imputation of incomplete categorical variables", *Computational Statistics and Data Analysis*, 54: 2267-2275

[32] White, I.R., Wood, A. and Royston, P. (eds), 2007, "Editorial: Multiple imputation in practise", *Statistical Methods in Medical Research*, 16: 195-197

[33] White, I.R., Royston, P. and Wood, A., 2011, "Multiple imputation using chained equations: Issues and guidance for practise", *Statistics in Medicine*, 30: 377-399

[34] Wittenberg, M., 2008, *"Nonparametric estimation when income is reported in bands and at points"*, Cape Town: Economic Research Southern Africa Working Paper Number 94

About DatatFirst

DataFirst is a research unit at the University of Cape Town engaged in promoting the long term preservation and reuse of data from African Socioeconomic surveys. This includes:

- the development and use of appropriate software for data curation to support the use of data for purposes beyond those of initial survey projects
- liaison with data producers - governments and research institutions - for the provision of data for reanalysis
- research to improve the quality of African survey data
- training of African data managers for better data curation on the continent
- training of data users to advance quantitative skills in the region.

The above strategies support a well-resourced research-policy interface in South Africa, where data reuse by policy analysts in academia serves to refine inputs to government planning.

DataFirst

www.datafirst.uct.ac.za
Level 3, School of Economics Building, Middle Campus, University of Cape Town
Private Bag, Rondebosch 7701, Cape Town, South Africa
Tel: +27 (0)21 650 5708

University of Cape Town • iYunivesithi yaseKapa • Universiteit van Kaapstad •

SPES BONA