

# **‘The impact of multiple imputation of coarsened data on estimates of the working poor in South Africa’**

**Claire Vermaak<sup>1</sup>**

## **Abstract**

South African household surveys typically contain coarsened earnings data, which consist of a mixture of missing earnings values, point responses and interval responses. This paper uses sequential regression multivariate imputation to impute missing and interval-censored values in the 2000 and 2006 Labour Force Surveys, and compares poverty estimates obtained under several different methods of reconciling coarsened earnings data. Estimates of poverty amongst the employed are found not to be sensitive to the use of the multiple imputation approach, but are sensitive to the treatment of workers reporting zero earnings. Including workers who plausibly report zero earnings, the proportion of workers earning less than R500 per month falls by almost a third between 2000 and 2006.

***JEL Classification:* C81, J31, I32**

***Keywords:* coarsened data; multiple imputation; poverty; wage distribution; working poor**

***DRAFT VERSION: Please do not cite***

---

<sup>1</sup> Lecturer, School of Economics and Finance, University of KwaZulu-Natal, Durban, South Africa.  
[vermaak@ukzn.ac.za](mailto:vermaak@ukzn.ac.za)

## **1 Introduction**

Household surveys usually contain earnings data that are coarsened, in that some earnings values are missing through item non-response, while earnings responses consist of both point and interval values. This makes it difficult to construct a continuous earnings variable with which to analyse poverty and inequality. Empirical studies on poverty and inequality in South Africa typically ignore the missing data, and combine point observations with interval midpoints to create a single earnings variable. However, both of these approaches are problematic. By ignoring missing data, researchers implicitly assume that the data are missing completely at random; if they are not, the resulting estimates will be biased. Second, using interval midpoints ignores the distribution of earnings within intervals; any subsequent distribution-based estimates will thus be biased.

Imputation provides a means of utilising data that are subject to item non-response, by assigning a plausible value to missing data. In addition, multiple imputation techniques enable the researcher to generate standard errors that properly reflect the uncertainty involved in the imputation process. Using this methodology thus enables the researcher to construct a continuous earnings variable from coarsened data, and to use it to analyse poverty levels and trends, while acknowledging the additional uncertainty arising from the use of imputed data.

The remainder of the paper is structured as follows. The next section briefly reviews the literature on data coarsening, outlining how it affects earnings estimates. Section 3 considers the methodology of multiple imputation, and how it differs from the usual methods applied to coarsened earnings data in South African household surveys. Section 4 presents the data used in the analysis. Section 5 presents the empirical estimates of rates of poverty amongst the employed. In particular, this section compares the poverty estimates that result from using several different methods of dealing with coarsened data. Section 6 presents some descriptive statistics of levels and trends in poverty amongst the employed, using the multiply-imputed datasets. Finally, section 7 concludes and explores the implication of these findings for the estimation of poverty in South Africa.

## **2 The problem of missing and coarsened data**

Survey data are frequently incomplete, in that some of the observational units comprising the sample do not respond to one or more of the parts of the questionnaire. When the available (observed) data are analysed as if they make up the complete sample, researchers implicitly ignore the mechanism which created the missing data. In addition to decreased precision that results from analysing a smaller dataset, resulting inferences may also be biased if the observed data differ systematically from the unobserved data.

There are several different ways that non-response for a variable can be generated, as categorised by Rubin (1987). The data are said to be missing completely at random (MCAR) if the missingness depends on neither the observed nor the unobserved (missing) data. The missing data on a particular variable thus constitute a simple random sample of that variable. If the missingness depends on the observed data, but not on the unobserved data, then the data are said to be missing at random (MAR). Under both MCAR and MAR, the missing-data structure is ignorable, since inferences can be drawn on parameters of interest without knowing the nature of the missingness mechanism.

If the missingness depends on both the observed and unobserved data, such that the probability of a value being missing depends on the unobserved value itself, even after conditioning on the observed values, then the data are said to be missing not at random (MNAR). In such cases, the missingness mechanism is non-ignorable, in that it must be taken into account when drawing inferences on parameters of interest.

If the data are MCAR, analysis of the observed data will produce unbiased estimates of parameters of interest, but there will be some loss of precision in accordance with the smaller sample size. However, MCAR is extremely unlikely in practice (Durrant, 2005). If the observed data are analysed as if they comprise the complete dataset when the data are MAR or MNAR, resulting parameter

estimates may be biased substantially. The extent of the bias is a function of the fraction of missing data (Lacerda *et al*, 2008: 61).

In addition to data that are entirely missing, data coarsening is also common in surveys. Data are said to be coarsened when they contain some combination of point (actual) responses, interval (bracket) responses and missing values (item non-response). Data on income, assets and earnings in household surveys are often coarsened because survey instruments provide bracketed response options in order to reduce information that would otherwise be lost through item non-response (Heeringa *et al*, 1997). However, such data are complex for researchers to work with, as it is difficult to combine the different types of data values into a single monetary measure of wellbeing. The mechanism which generates the data coarsening has similar properties to the missingness mechanism; if data are coarsened at random (CAR), then the mechanism which generates the interval censoring and the missing data is ignorable (Heitjan and Rubin, 1991). However, unless the data are coarsened completely at random, analysing only the uncoarsened portion of the data will result in biased parameter estimates.

### 3 Imputation methodology

Imputation is the process by which missing data are filled in using plausible values, so that techniques developed for analysing complete datasets can be used. Single imputation involves replacing each missing value with a single predicted value, to create a single complete dataset. Examples of single imputation methods include mean substitution, regression imputation and hotdeck imputation<sup>2</sup>. However, the fundamental flaw underlying single imputation techniques is that they fail to take into account that imputed values are more uncertain than observed values. Thus the standard errors of any estimates that are subsequently obtained from the singly imputed dataset are likely to be understated, in that they do not reflect this additional uncertainty (Rubin, 1987).

Multiple imputation involves applying a stochastic imputation model to the missing data problem. The model is applied  $m$  times, creating  $m$  plausible datasets, and thus multiple imputation produces a distribution of imputed values which reflects the uncertainty involved in the imputation process. Estimates of interest obtained separately from each of the  $m$  imputed datasets are then combined as follows, using Rubin's rules (Rubin, 1987). Let  $Q_i$  represent the estimate of interest from the  $i^{\text{th}}$  imputed dataset, and let  $U_i$  represent the variance of that estimate. Then the overall combined point estimate is given by

$$\bar{Q} = \sum_{i=1}^m Q_i / m$$

and the variance of the combined estimate is given by

$$T = U + (1 + m^{-1})B$$

where  $U = \sum_{i=1}^m U_i / m$  is the average within-imputation variance and  $B = \sum_{i=1}^m (Q_i - \bar{Q})^2 / (m-1)$

is the between-imputation variance. For large samples, the estimate of  $\bar{Q} \pm 1.96\sqrt{T}$  provides a 95 percent confidence interval for  $Q$ .

This paper uses a particular multiple imputation technique developed by Raghunathan *et al* (2001) for imputing missing values within a complex data structure, when the data are MAR. Called sequential regression multivariate imputation (SRMI), the method can be used to impute both data values that are entirely missing, and those that are known to be located within a particular interval. The method is used not only to impute coarsened earnings data, but also simultaneously imputes missing values of other variables that will be used in later analysis.

The SRMI method proceeds as follows. The variables to be used in the imputation model are ordered from the least to the most amount of missing values. Let the matrix  $\mathbf{X}$  represent all variables that are fully observed, while  $Y_1, \dots, Y_k$  represent the ordered variables that contain missing values. The first

---

<sup>2</sup> For a review of imputation techniques, see for example, Durrant (2005) and Lacerda *et al* (2008).

imputation begins by regressing  $Y_1$  on  $\mathbf{X}$ , and imputing values for  $Y_1$  using random draws from the appropriate predictive distribution for  $Y_1$ . For example, a normal linear regression model is used when  $Y_i$  is a continuous variable, a logistic model when  $Y_i$  is binary, and a polytomous logit model when  $Y_i$  is categorical. An interval regression model is used to impute values for variables containing both missing and interval values, following a truncated normal distribution when interval values are reported, and a normal distribution without bounds when values are missing.

Since its missing values have now been imputed,  $Y_1$  is appended to the set of predictor variables. Thus  $Y_2$  is now regressed on  $\mathbf{X}$  and the imputed  $Y_1$ , and values are imputed for  $Y_2$ , and so on until all  $Y$  variables have been imputed using all previously imputed variables as covariates. The imputation process is then repeated, updating the regression parameters  $\theta$  with parameters drawn from the now-complete distribution. This cycle is repeated until the imputed values and parameters converge to a stable distribution. This produces the first imputed dataset. The entire procedure is then repeated  $m$  times, to produce  $m$  imputed complete datasets. Estimates of interest, and their standard errors, are produced using Rubin's rules, as outlined above.

#### 4 Data

This study makes use of data collected by the Labour Force Surveys (LFSs) of September 2000 and September 2006. The LFSs are chosen because of the consistency of the survey instrument in collecting labour market information across time. Using the 2006 survey allows recent estimates of poverty to be made, while using the 2000 survey allows for a sufficient time period over which to assess trends in poverty. Additionally, this time period encompasses a number of important legislative developments which can be expected to have had an impact on the functioning of the labour market, and hence on poverty amongst the employed. In particular, as a result of the 2002 amendment to the Basic Conditions of Employment Act, minimum wage determinations were extended to a number of sectors in which workers traditionally have been vulnerable (such as domestic work and agricultural wage employment). Therefore the extent of poverty among the employed, and particularly the wage employed, can be expected to have declined over this time.

The international literature generally defines the working poor as “those who work *and* who belong to poor households” (Majid, 2001: 2; emphasis in original). However, by identifying poverty at the household level, this definition conflates the earnings of the individual worker with the earned and non-earned income of other members of the household. Changes in the poverty status of an individual worker may then result from changes in his/her individual earnings, changes in the income of other household members, or changes in the composition of the household. Given the substantial increases in social transfers and the changes in household dynamics in South Africa over the study period, the effectiveness of the labour market in redistributing income to the bottom tail of the earnings distribution would be obscured by such a definition.

This study instead defines the working poor as those individuals who work but whose earnings are insufficient to lift them above an individually-defined poverty line. The advantage of such a definition of the working poor is that it enables an analysis of how interactions between the labour market and the characteristics of the individual relate to his/her poverty status at different points in time. The disadvantage of using this definition is that it does not consider income-pooling within the household, and thus can say nothing about how the poverty status of households is affected by changes in the earnings of individual members. Therefore this study in fact amounts to a study of low-earning workers, rather than a general study of poverty.

This study uses two poverty lines, in order to assess the effects of imputation of coarsened data on differently-specified poverty lines, and to assess the extent of changes in poverty at different points in the earnings distribution. The first poverty line is set at R150 per month at real 2000 prices. This poverty line corresponds approximately in 2006 to the boundary between the second (R1 – R200) and

third (R201 – R500) earnings brackets in the LFSs, when the brackets are converted into real terms<sup>3</sup>. Although this poverty line has been chosen for its relationship with the earnings bracket, it is close in value to the \$2 per day international poverty line, which amounts to R159 in 2000 prices<sup>4</sup>.

The second poverty line is set at R500 per month, at real 2000 prices. This poverty line corresponds to a value slightly below the midpoint of the fourth earnings bracket (R501 – R1000) in 2006, when converted into real terms. This poverty line represents an earnings value approximately 25 percent higher than the household subsistence level per adult equivalent (Potgieter, 1999) in 2000 prices.

The extent to which earnings data are coarsened in the September 2000 and 2006 LFSs is illustrated in Table 1. While most individuals have earnings reported as a point figure (that is, a single numerical value), the proportion of workers with such an earnings value falls between the two surveys, and a growing proportion of workers report their earnings as a bracket figure only. In addition, a substantial but decreasing proportion of workers report that they have zero earnings, while a growing proportion of workers report no earnings information. Analysing only workers with (positive) point earnings information would imply that other information from 22 percent of workers in 2000, and 33 percent of workers in 2006, would be ignored. Even if just workers with zero and missing earnings information are excluded from the analysis, more than 10 percent of the sample of workers is lost. Thus implementing methods for dealing with coarsened data enables a much greater proportion of the data to be analysed than would otherwise be possible.

**Table 1. Type of earnings value reported by the employed**

<b>Proportion of all employed:</b>	<b>2000</b>	<b>2006</b>
Point response	0.776 (0.006)	0.667 (0.011)
Bracket response	0.107 (0.005)	0.231 (0.010)
Zero earnings	0.079 (0.004)	0.035 (0.005)
Missing (includes responses 'Don't know' and 'Refuse')	0.038 (0.002)	0.067 (0.006)

Source: LFS September 2000 and 2006

Notes: Standard errors in parentheses

All estimates are weighted to population levels using weights provided by StatsSA

## **5 Poverty estimates, by method of estimation**

There has been considerable research and debate on levels and trends in poverty and inequality in post-apartheid South Africa. The issue of whether poverty and inequality have increased or decreased since the advent of democracy is of great importance, since it goes to the heart of the effectiveness of government's social and economic policies. The emergence of nationally representative household surveys as a data source from 1993 onwards provided researchers with a wide variety of data with which to conduct such studies. Although there is considerable variation in the data sets used, including the Census (Ardington *et al*, 2006; Leibbrandt *et al*, 2006), October Household Surveys (Meth and Dias, 2004; Leibbrandt *et al*, 2005), Income and Expenditure Surveys (Leibbrandt *et al*, 2005; Hooegeveen and Özler, 2006), Labour Force Surveys (Meth and Dias, 2004; Leibbrandt *et al*, 2005) and, more recently, the All Media and Products Surveys (van der Berg *et al*, 2006; van der Berg *et al*, 2008), most researchers use household income, comprising both earned and unearned income, or household expenditure, as a measure of wellbeing. Most authors focus on the 1995 to 2000 period, and find that inequality rises, but that the direction and extent of any change in poverty is dependent

<sup>3</sup> Using the average CPI for metropolitan areas for 2006 of 1.34 (Statistics South Africa, 2007)

<sup>4</sup> Converted from US dollars to South African Rands using purchasing power parity of \$1 = R2.65 in 2000.

on the poverty line used. For the more recent period, van der Berg *et al* (2008) find that poverty, based on per capita income data from the AMPS, has decreased since 2000.

Amongst these studies, Ardington *et al* (2006) is the only one which multiply imputes missing income values. They work mainly with the Census 2001 data, and find that income data are missing for 16 percent of individuals, while a large (but unspecified) proportion of individuals have zero recorded incomes. They therefore apply the SRMI technique to impute income values for these individuals, and then sum individual income across each household and divide by household size, in order to analyse income per capita. They find that SRMI methods produce higher estimates of mean per capita income, and lower estimates of poverty rates, than without using imputation. However, income values in the Census are collected only in brackets. For the majority of their paper, the authors assign each individual the midpoint of their income bracket as their point income. Although they then test the sensitivity of their estimates of poverty and inequality to this approach, they do not do so by applying interval regression SRMI. Rather, they distribute income within each bracket according to the empirical distribution of individual income from the Income and Expenditure Survey conducted in the same year. However, since the IES data are only collected at five-yearly intervals, this technique is not applicable to the LFS data used in the present study. Ardington *et al* (2006) find that their estimates are not very sensitive to the method applied to incomes reported in brackets. Overall, they find that poverty and inequality rise between 1996 and 2001, which confirms the results of other literature which uses these datasets without perform multiple imputation.

In contrast to the wealth of studies using income or expenditure data, relatively few studies focus specifically on the role of earnings in changes in poverty or inequality. Leite *et al* (2006) analyse trends in earnings inequality, but not poverty, up to 2004, while Cichello *et al* (2001; 2005) analyse earnings dynamics using panel data, but only amongst Africans in KwaZulu-Natal and not focusing specifically on the working poor. Estimates of the number of the working poor at particular points in time are contained in Casale *et al* (2004) and in Posel and Casale (2005) as part of a wider study of other issues. They find that the number of the employed who fall below a poverty line of \$2 a day in real terms (2000 prices) more than doubles between 1995 and 2003. The present study therefore investigates more thoroughly how the incidence of poverty amongst the employed has changed since 2000, and briefly examines some factors that may contribute to the employed being poor.

The estimates of poverty amongst the employed in this section are presented in two broad categories. In the first category, all missing earnings values are excluded, and interval midpoints are used as estimates of earnings for those who report their earnings in a bracket. In the second category, a multiple imputation approach is used progressively to produce estimates of interval, missing and zero earnings values.

In order to impute earnings values, SRMI was carried out including standard earnings equation covariates in the imputation model<sup>5</sup>. Thus missing values for variables such as age, working hours and education were imputed as part of the process of imputing earnings. Of particular interest for this study, the natural logarithm of monthly earnings was imputed using interval regression, in order to deal simultaneously with point observations, interval-censored observations, right-censored observations and missing observations.

Table 2 outlines the approaches within each of the two categories in terms of how interval, missing and zero earnings responses are treated, and the effect of each approach on the sample size and on the mean of the natural logarithm of earnings<sup>6</sup>, for the LFS 2006. The sample size of the employed, when only point and interval earnings responses are considered (approach A), is 24 097. Including all workers who report zero earnings (B) increases the sample size to 25 567, and decreases mean

---

<sup>5</sup> Multiple imputation was implemented in Stata using the add-on function *ice*, with each of the five multiply-imputed datasets being produced using ten cycles. The resulting multiply-imputed datasets were analysed using the add-on function *mim*.

<sup>6</sup> Earnings are imputed in logarithmic form, therefore the mean is also presented in log-form.

earnings. Excluding workers for whom zero earnings are implausible (C) lowers the sample size, and raises mean earnings, slightly. Using SRMI interval regression to impute interval values (D), rather than using interval midpoints (A), makes little difference to mean earnings, but imputing earnings for workers with missing earnings data (E) raises both the mean and the sample size. The sample size reaches its maximum when both zero earnings and imputed values for missing earnings are included. Treating all zero responses as genuine (F), all as values to be imputed (G) or according to their plausibility (H) affects mean earnings substantially, but the full sample size is maintained in each case. Table 1 suggests that the way in which workers who report zero earnings are treated by the study has a much larger effect on the earnings distribution than the imputation of missing and interval-censored earnings data.

**Table 2. Approaches to the treatment of different types of earnings responses**

		Treatment of earnings responses			Sample size	Mean of ln(earnings)
		Interval responses	Missing responses	Zero responses		
<b>Approaches without imputation</b>	A	Midpoints	Omitted	Omitted	24 097	7.308 (0.044)
	B	Midpoints	Omitted	All included	25 567	6.999 (0.076)
	C	Midpoints	Omitted	Plausible zeroes included	25 502	7.012 (0.074)
<b>Approaches using SRMI</b>	D	Imputed	Omitted	Omitted	24 097	7.304 (0.045)
	E	Imputed	Imputed	Omitted	25 294	7.347 (0.047)
	F	Imputed	Imputed	All included	26 764	7.056 (0.077)
	G	Imputed	Imputed	All imputed	26 764	7.293 (0.050)
	H	Imputed	Imputed	Plausible zeroes included; implausible zeroes imputed	26 764	7.081 (0.074)

Source of estimates: LFS September 2006

Notes: Standard errors in parentheses

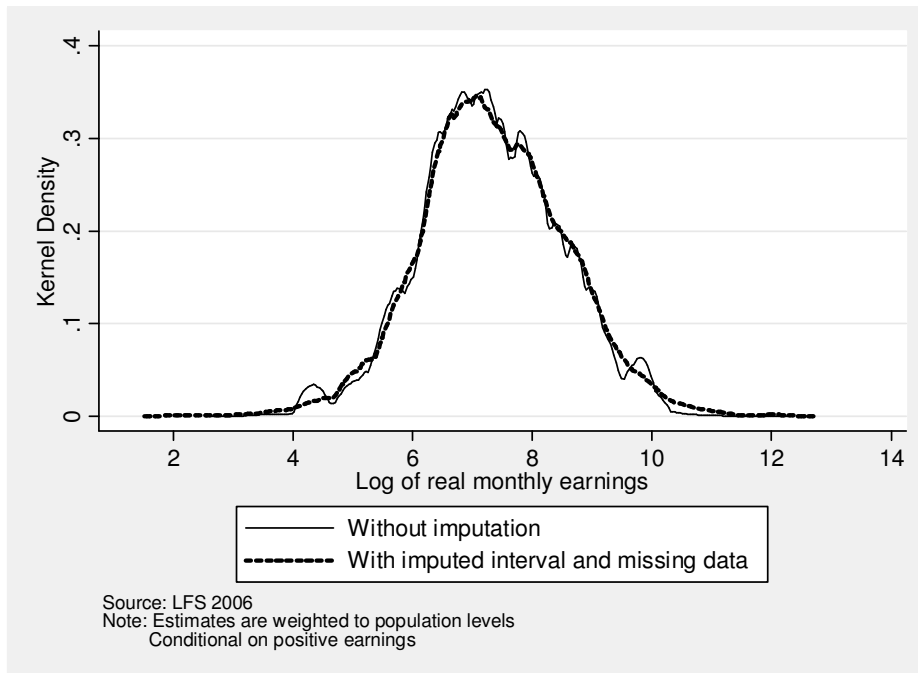
Estimates of mean earnings are weighted to population levels using weights provided by StatsSA

The distribution of log real monthly earnings in the 2006 LFS, and the effects of multiple imputation on this distribution, are presented in the kernel density estimates in figures 1 and 2<sup>7</sup>. Without using SRMI, the kernel exhibits ‘bumps’ representing the allocation of the earnings value at the midpoint of the interval to workers who report their earnings in a bracket. For example, the natural logarithm of the midpoint of the R1 – R200 earnings bracket, converted into real terms, is 4.3, which is the location of the first ‘bump’. The main effect of the SRMI for interval and missing data is thus to smooth the kernel, by applying a truncated normal distribution to interval-censored earnings values.

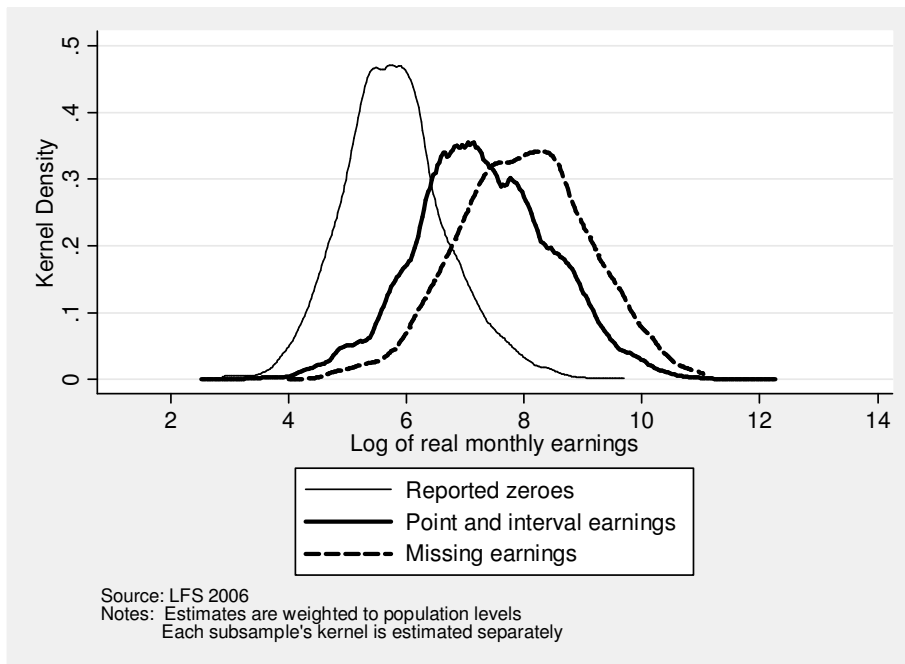
The application of SRMI produces quite different distributions of earnings for workers who do not report earnings, workers who report zero earnings, and workers who report interval-censored or point earnings. Figure 2 illustrates that imputed values for workers who report zero earnings are substantially lower, and less widely dispersed, than imputed values for other workers, although the imputed values nevertheless lie considerably above zero. Imputed earnings values are highest for workers with missing earnings information, which is consistent with the finding of other authors that workers who do not report their earnings are older, more educated, and more likely to be white and living in an urban area, all of which are characteristics that are associated with higher earnings values (Posel and Casale, 2005).

<sup>7</sup> All density estimates use an Epanechnikov kernel, and the Silverman (1986) rule-of-thumb bandwidth selector.

**Figure 1. The distribution of earnings, without and with imputation, in 2006**



**Figure 2. The distribution of imputed zero, missing and interval-censored earnings, in 2006**



### 5.1 Poverty estimates, without using imputation

The most common method used by researchers to reconcile point and interval earnings data in South African household surveys is to assign interval respondents an earnings value equal to the midpoint of the interval (*cf.* Leibbrandt *et al*, 2006; Leite *et al*, 2006). This method is used in this paper for all three of the approaches that do not use multiple imputation. It is not possible to use the empirical intra-band allocation approach of Ardington *et al* (2006) since the IES data are only collected at five-yearly intervals, and are thus not compatible with the biannually-collected LFS data.

However, there is a further issue to consider when constructing an earnings variable, in that a substantial proportion of the employed report that they earn zero income. As shown in Table 1, the proportion of all workers who report non-zero working hours but zero earnings is 7.9 percent in 2000 and 3.5 percent in 2006. Since the Labour Force Survey questionnaire asks respondents to report their total salary at their main job, but does not include a prompt for payments in-kind, individuals who do not receive a cash wage, such as those engaged in subsistence agriculture or working without pay in a family business, are likely to report zero earnings. The level of scepticism with which workers reporting non-zero working hours but zero earnings are treated, and hence whether such zero earnings values are included in the analysis, makes a substantial difference to estimates of poverty.

This study takes three approaches to the treatment of zero earnings. The first (approach A) is to condition the estimates on positive earnings being reported, thus excluding all individuals who report zero earnings. This is the most common approach used by researchers working with the October Household Surveys and Labour Force Surveys. In the second approach (B), all reported zero earnings values are treated as being the genuine earnings of those individuals. In the third approach (C), individuals who report zero earnings are included in the poverty estimates only if it is regarded as ‘plausible’ that they earn no cash wage. In this approach, earning a wage of zero is considered plausible if, when answering the question “In the last seven days, did ... do any of the following activities, even for only one hour?”, the individual only performed unpaid tasks, such as working in a household business or in subsistence agriculture<sup>8</sup>.

Table 3 shows the results of these three approaches. Conditional on positive earnings, 335 000 workers earn less than R150 per month (in real 2000 prices) in 2006, amounting to 2.9 percent of all workers. 1.8 million individuals, or 15.7 percent of workers, earn less than R500 per month. These estimates can be regarded as a lower bound for poverty at each poverty line, since they exclude all workers reporting zero earnings. When all such workers are included, the number of the working poor rises by 510 000, resulting in the poverty rate rising to seven percent at the lower poverty line, and 19.3 percent at the upper line. These estimates can be regarded as an upper bound for poverty at each poverty line, since they include as poor all workers reporting zero earnings. Since most reports of zero earnings can be regarded as plausible, including only workers with plausible zero earnings produces estimates that are very similar to the upper bound.

**Table 3. Poverty amongst the employed estimated without using imputation, by method of treatment of zero earnings**

	<b>Approach</b>		
	A (positive earnings only)	B (incl. all zeroes)	C (incl. plausible zeroes)
<b>Poverty Line 1: R150 per month</b>			
Working poor ('000s)	335 (66)	845 (190)	822 (184)
Headcount ratio	0.029 (0.002)	0.070 (0.007)	0.068 (0.007)
<b>Poverty Line 2: R500 per month</b>			
Working poor ('000s)	1 815 (332)	2 325 (455)	2 302 (449)
Headcount ratio	0.157 (0.010)	0.193 (0.014)	0.191 (0.013)

Source: LFS September 2006

Notes: Poverty lines in real 2000 prices

Standard errors in parentheses

All estimates are weighted to population levels using weights provided by StatsSA

<sup>8</sup> Specifically, individuals with zero earnings who gave only responses d), e) or f) to question 2.1

## 5.2 Multiple imputation of interval and missing earnings values

In this section, sequential regression multiple imputation (SRMI) is used to impute earnings values. Throughout this section, interval regression is used to impute earnings values for the bracket responses, but several different approaches are used for individuals with missing or zero earnings. This allows for comparison with the unimputed estimates. Table 4 presents the estimates of poverty rates, conditional on positive earnings being reported. In the first column of results (A), the estimates without using imputation from Table 3 are repeated, for comparison purposes. In the second column (D), earnings values are imputed for the bracket responses, but not for missing earnings. In the third column (E), earnings values are imputed for both the bracket and missing earnings responses.

**Table 4. Poverty amongst the employed, by extent of multiple imputation**

	<b>Approach</b>		
	A	D	E
	(without imputation; midpoints for intervals)	(imputation for intervals only)	(imputation for intervals and missing data)
<b>Poverty Line 1: R150 per month</b>			
Working poor ('000s)	335 (66)	335 (66)	367 (72)
Headcount ratio	0.029 (0.002)	0.029 (0.003)	0.030 (0.003)
<b>Poverty Line 2: R500 per month</b>			
Working poor ('000s)	1 815 (332)	1 891 (347)	2 007 (365)
Headcount ratio	0.157 (0.010)	0.164 (0.011)	0.162 (0.011)

Source: LFS September 2006

Notes: Poverty lines in real 2000 prices  
Standard errors in parentheses

All estimates are conditional on positive earnings being reported and are weighted to population levels using weights provided by StatsSA

Table 4 thus compares the imputation of interval and missing earnings values, to the use of interval midpoints. Since the poverty line of R150 per month corresponds to the boundary between the second and third earnings brackets, imputing values for the intervals (approach D) produces exactly the same estimate of the number and rate of poverty as using the interval midpoints (approach B). However, using the midpoint method assigns the same value to everyone in a bracket, while using interval regression imputation produces a truncated normal distribution within the bracket. Thus the estimate of the depth of poverty would differ by technique, although the poverty headcount does not. There is a difference in estimates between the two techniques at the R500 poverty line, since this line intersects the fourth earnings bracket. The midpoint of this bracket, R560, is greater than the poverty line, thus all individuals reporting earnings in the fourth bracket are classified as non-poor by the midpoint technique. Using imputation, some individuals from within this bracket are classified as poor and others as non-poor. Thus the poverty rate estimated using the imputation technique is slightly higher, at 16.4 percent, than using the midpoint technique, at 15.7 percent. The extent to which poverty estimates are affected by using interval regression, rather than bracket midpoints, to impute interval responses thus depends on the extent to which the poverty line bisects an earnings bracket.

Approach E presents poverty estimates when both interval and missing earnings values are imputed. Approximately 32 000 of the 848 000 workers with missing earnings data are classified at the very lower end of the imputed earnings distribution, while a further 84 000 workers earn between R150 and R500 per month. Thus excluding workers with missing earnings data by using the non-imputation approach under-estimates the poverty rate by 0.1 percentage points at the lower poverty line, and by 0.5 percentage points at the upper poverty line.

One of the major contributions of multiple imputation methods, compared to single imputation methods, is that it provides standard errors that properly reflect all sources of uncertainty in the calculation of estimates. Thus although the sample size is larger for the imputation approaches than the non-imputation approaches, the standard errors are also larger, reflecting variability amongst the imputations. Although the approaches that use SRMI produce larger estimates of poverty amongst the employed than the non-imputation methodology, none of the estimates differ by more than one standard error. Thus, conditional on positive earnings being reported, multiple imputation of interval-censored and missing earnings data does not produce significantly different estimates of poverty amongst the employed than the traditional non-imputation methodology.

In Table 5, estimates of poverty amongst the employed are presented in which SRMI is again used for both interval and missing earnings data. However, the estimates now differ according to the treatment of zero reported earnings. Imputing earnings values for all workers who report zero earnings (approach G), rather than taking all such earnings values at face value (F), roughly halves the number and proportion of workers who earn less than R150 per month. Imputing values only for workers who implausibly report zero earnings results in an estimated 6.3 percent of workers earning less than R150 per month, and 18.7 percent earning less than R500 per month. These estimates are quantitatively similar to the figures of 6.8 and 19.1 percent respectively at the two poverty lines, produced without using imputation (approach C). Once again, none of the SRMI estimates is significantly different to its corresponding non-imputation estimate.

**Table 5. Poverty amongst the employed, by method of imputation of reported zero earnings**

	<b>Approach</b>		
	F	G	H
	(all zeroes included)	(all zeroes imputed)	(implausible zeroes imputed)
<b>Poverty Line 1: R150 per month</b>			
Working poor ('000s)	877 (195)	425 (93)	815 (182)
Headcount ratio	0.068 (0.007)	0.033 (0.003)	0.063 (0.007)
<b>Poverty Line 2: R500 per month</b>			
Working poor ('000s)	2 517 (486)	2 281 (445)	2 415 (470)
Headcount ratio	0.195 (0.014)	0.177 (0.014)	0.187 (0.014)

Source: LFS September 2006

Notes: Poverty lines in real 2000 prices  
Standard errors in parentheses

All estimates are weighted to population levels using weights provided by StatsSA

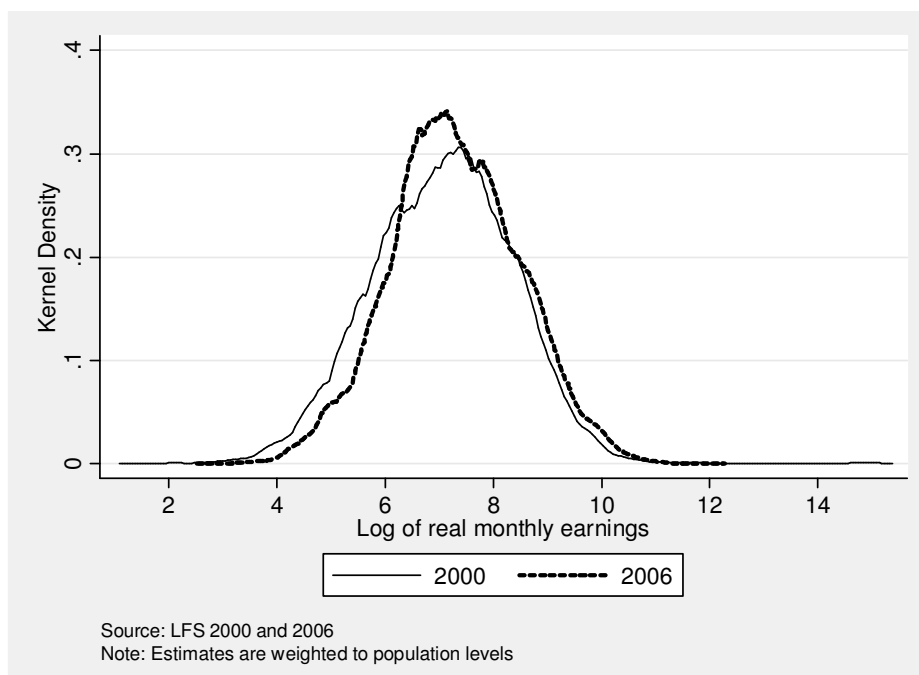
Which of the approaches presented above produces the 'right' poverty rate at a given poverty line? It depends largely on what the researcher believes about the validity of zero earnings values. What is important is that the way in which workers who report zero-earnings are treated is consistent with how workers who report positive earnings are treated. Since workers are not prompted to report in-kind earnings, the value of reported earnings understates total remuneration for *all* workers who receive in-kind benefits, regardless of whether or not they also receive cash wages. Thus, for the remainder of this study, workers who plausibly report zero earnings are included in the poverty analysis, since reported earnings consists only of cash earnings for workers reporting positive earnings too. However, although estimated *levels* of poverty are different if zero earnings are treated differently, the direction of poverty *trends* is robust to the method of treatment of zero earnings.

## 6 Trends in poverty amongst the employed

There have been substantial changes in the legislative framework of the South African labour market since the end of apartheid, aimed at setting minimum employment standards, regulating organised bargaining and redressing discrimination. As a result, poverty amongst the employed, and particularly the wage-employed, can be expected to have declined. However, labour market trends over this period may have acted to distribute such gains unevenly amongst the employed. The feminisation of the labour force, growing unemployment, informalisation of work and growth in self-employment suggest that some types of workers may be crowded into self-employment or jobs in the informal sector which are not covered by the new legislation.

The kernel density estimate in figure 3 supports these conjectures. There is an unambiguous improvement in real earnings between 2000 and 2006 for those at the bottom of the earnings distribution. However, the log earnings distribution also narrows over time, such that lesser improvements in earnings are achieved higher up in the distribution.

*Figure 3. The distribution of real log monthly earnings, 2000 and 2006*



Poverty rates amongst the employed estimated at different poverty lines further illustrate these changes. Table 6 below presents estimates of the number and proportion of workers earning less than R150 and R500 per month respectively. Approximately 1.3 million workers earn less than R150 per month in 2000, with an additional 2 million earning between R150 and R500 per month. However, the rate of poverty amongst the employed declines substantially between 2000 and 2006. Both the absolute number and the proportion of low-earning workers declines between the two years, by more than 40 percent at the lower of the two poverty lines, but to a smaller extent at the higher poverty line. Thus, on aggregate, workers are better off in 2006 than they were in 2000, but the improvement is larger at the very bottom of the earnings distribution than it is higher up in the distribution.

Part of the observed decrease in poverty is attributable to the decline in the reporting of zero earnings between 2000 and 2006. Conditional on positive earnings being reported, the decrease in the poverty rate is approximately half the size of that reported in table 6. However, the poverty rate amongst the employed is nonetheless significantly lower in 2006 than it was in 2000.

**Table 6. Poverty levels and trends, 2000 and 2006**

<b>Poverty Line 1: R150 per month</b>		<b>2000</b>	<b>2006</b>	<b>Change (%)</b>
Working poor ('000s)		1 387 (65)	815 (182)	-41.2
Headcount ratio		0.114 (0.005)	0.063 (0.007)	-44.4
<b>Poverty Line 2: R500 per month</b>				
Working poor ('000s)		3 304 (89)	2 415 (470)	-26.9
Headcount ratio		0.271 (0.007)	0.187 (0.014)	-30.8

Source: LFS September 2000 and 2006

Notes: Poverty lines in real 2000 prices

Standard errors in parentheses

All estimates are weighted to population levels using weights provided by StatsSA

Differences in the decline in poverty at different points in the earnings distribution suggest that gains may be unevenly distributed amongst different groups of the employed. This is indeed the case, as is summarised here across two dimensions, namely education and sector of employment. The descriptive statistics in the remainder of the paper use the R500 per month poverty line, as this line captures a larger number of low-earning workers than does the lower poverty line.

Education can be expected to offer protection against low earnings, and this is reflected by the proportion of workers who earn less than R500 per month being lower at each successively higher level of education, illustrated in table 7. In addition, there is a substantially greater decline in the poverty rate between 2000 and 2006 amongst those workers who have completed Matric, than amongst workers with less than a Matric education.

**Table 7. Poverty amongst the employed, by highest level of education completed**

<b>Education level</b>		<b>2000</b>	<b>2006</b>	<b>Change (%)</b>
None	Number ('000s)	656 (28)	374 (86)	-43.1
	Headcount ratio	0.597 (0.015)	0.507 (0.022)	-15.0
Grade 1-7	Number ('000s)	1 381 (45)	881 (188)	-36.2
	Headcount ratio	0.436 (0.010)	0.364 (0.017)	-16.6
Grade 8-11	Number ('000s)	934 (33)	866 (162)	-7.3
	Headcount ratio	0.258 (0.008)	0.214 (0.015)	-17.2
Matric	Number ('000s)	275 (21)	260 (46)	-5.5
	Headcount ratio	0.117 (0.008)	0.073 (0.006)	-37.5
Diploma/Degree	Number ('000s)	57 (7)	35 (8)	-39.1
	Headcount ratio	0.029 (0.003)	0.016 (0.003)	-44.0

Source: LFS September 2000 and 2006

Notes: Poverty lines in real 2000 prices

Standard errors in parentheses

All estimates are weighted to population levels using weights provided by StatsSA

However, these changes occur against a backdrop of rising average education levels amongst the employed as a whole. Thus education, and secondary schooling in particular, in fact offers *less* protection against low-earning work in 2006 than it did in 2000. While more than 60 percent of low earning workers have no schooling or primary education only in 2000, the proportion of low earning workers who have an education level of grade eight or higher rises by ten percentage points during this six year period.

Poverty rates would also be expected to vary substantially according to the sector in which the employed work. Less than 20 percent of wage-employed workers earn less than R500 per month in 2000, while more than 50 percent of self-employed workers do so. In addition, the poverty rate falls to a greater extent amongst the wage-employed than amongst the self-employed, such that the proportion of low earning workers who are self-employed rises from 46 percent in 2000 to 49 percent in 2006. The poverty rate drops by almost a half amongst workers who can be expected to benefit most from improvements in labour legislation: wage-employed workers in the formal sector. In contrast, the drop in the poverty rate is smallest amongst self-employed workers in the informal sector, who thus make up a rising proportion of the working poor over time.

**Table 8. Poverty amongst the employed, by type and sector of employment**

<b>Employment type</b>		<b>2000</b>	<b>2006</b>	<b>Change (%)</b>
<b>All wage-employed</b>	Number ('000s)	1 785 (53)	1 244 (234)	-30.3
	Headcount ratio	0.190 (0.006)	0.121 (0.009)	-36.2
Formal sector	Number ('000s)	747 (34)	460 (95)	-38.4
	Headcount ratio	0.100 (0.005)	0.054 (0.006)	-45.5
Informal sector	Number ('000s)	1 038 (34)	784 (143)	-24.5
	Headcount ratio	0.539 (0.011)	0.430 (0.019)	-20.3
<b>All self-employed</b>	Number ('000s)	1 518 (67)	1 171 (244)	-22.9
	Headcount ratio	0.543 (0.014)	0.447 (0.025)	-17.7
Formal sector	Number ('000s)	80 (10)	45 (16)	-44.0
	Headcount ratio	0.133 (0.016)	0.068 (0.019)	-48.8
Informal sector	Number ('000s)	1 438 (65)	1 126 (234)	-21.7
	Headcount ratio	0.656 (0.013)	0.574 (0.025)	-12.4

Source: LFS September 2000 and 2006

Notes: Poverty lines in real 2000 prices

Standard errors in parentheses

All estimates are weighted to population levels using weights provided by StatsSA

## 7 Conclusion

South African household surveys, such as the Labour Force Surveys, contain coarsened earnings data, which consist of a mixture of missing earnings values, point responses and interval responses. The standard approach used by most researchers using these datasets is to create a continuous earnings variable by allocating interval midpoints to bracket respondents, while ignoring missing values.

However, such an approach will produce unbiased estimates only if the earnings data are coarsened at random, which is not the case.

In contrast, this study uses sequential regression multivariate imputation to produce multiple imputed datasets containing plausible values for both the missing and interval-reported earnings values. Compared to the standard approach, using SRMI significantly raises the estimate of mean earnings in the 2006 LFS, illustrating that the data were not coarsened at random. However, it does not significantly affect estimates of poverty amongst the employed. Imputed values for missing earnings observations mostly fall above the poverty line, while the imputation of interval responses affects estimates of poverty rates only to the extent that the poverty line bisects an interval.

This study goes on to show that the way in which workers who report earning zero income are treated in the analysis makes a far greater difference to estimates of poverty than does the treatment of missing and interval-reported data. Treating all reported zeroes as genuine, or imputing values only when reported zeroes seem implausible, produces significantly lower estimates of poverty than when earnings are imputed for all reported zeroes.

The study then turns to a brief assessment of trends in poverty amongst the employed between 2000 and 2006. The proportion of workers earning less than R150 per month falls from eleven to six percent during this time, but the improvement is smaller at a higher poverty line. In particular, education appears to offer less protection against poverty over time, while large improvements in earnings occur amongst those who would be expected to benefit most from changes in labour legislation over this period: wage-employed workers in the formal sector.

The analysis of low-earning workers presented here is merely suggestive, and many questions remain to be answered. What sorts of jobs generate such low monthly earnings? Are workers poor because their working hours are insufficient? Are low-earnings workers primary earners, or secondary earners, in their households? Do low-earning workers live in poor households? Thus although a specific focus on earnings is useful, because it enables an analysis of the effects of labour market trends and policies on poverty separate from the effects of the widely-documented extension of the social welfare system, it is also necessary to link low-earning workers with other sources of income in their households, in order to assess overall poverty outcomes.

In conclusion, multiple imputation certainly provides an attractive method of dealing with coarsened survey data. Provided that the imputation model is able to provide a plausible distribution of imputed values, this methodology can reduce non-response bias while also accounting for the additional variability that arises through imputation. However, implementing multiple imputation imposes costs on the researcher in terms of time and computing resources, both in creating and analysing the multiply imputed datasets. This study has shown that estimates of poverty amongst the employed are not significantly different when implementing SRMI than they are when ignoring missing data and assigning interval midpoints to interval respondents. Thus whether the benefits of the multiple imputation approach outweigh its costs, and whether this methodology becomes standard practice amongst poverty researchers as a result, remains to be seen.

## References

- Ardington, C., Lam, D., Leibbrandt, M. and Welch, M. (2006). "The sensitivity to key data imputations of recent estimates of income poverty and inequality in South Africa". *Economic Modelling* 23 (2006): 822– 835.
- Cichello, P.L., Fields, G.S. and Leibbrandt, M. (2001). "Are African Workers Getting Ahead in the New South Africa? Evidence from KwaZulu-Natal, 1993-1998". *Social Dynamics* 27(1): 120–139.
- Cichello, P.L., Fields, G.S. and Leibbrandt, M. (2005). "Earnings and Employment Dynamics for Africans in Post-apartheid South Africa: A Panel Study of KwaZulu-Natal". *Journal of African Economies*, 14(2): 143–190.
- Durrant, G.B. (2005). "Imputation Methods for Handling Item-Nonresponse in the Social Sciences: A Methodological Review". National Centre for Research Methods Working Paper Series, June 2005. National Centre for Research Methods and Southampton Statistical Sciences Research Institute, University of Southampton.
- Heeringa, S., Little, R.J.A., and Raghunathan, T. (1997). "Imputation of Multivariate Data on Household Net Worth". *Proceedings of the Survey Research Methods Section*, American Statistical Association 1997: 135-140.
- Heitjan, D.F., and Rubin, D.B. (1991). "Ignorability and coarse data". *The Annals of Statistics*, 19(4): 2244-2253.
- Hoogeveen, J.G. and Özler, B. (2006). "Poverty and Inequality in Post-Apartheid South Africa: 1995-2000". In Borhat, H. and Kanbur, R. (eds.) *Poverty and Policy in Post-Apartheid South Africa*. HRSC Press, Pretoria.
- Lacerda, M., Ardington, C. and Leibbrandt, M. (2008). "Sequential Regression Multiple Imputation for Incomplete Multivariate Data using Markov Chain Monte Carlo". Southern Africa Labour and Development Research Unit Working Paper Number 13. SALDRU, Cape Town.
- Leibbrandt, M., Levinsohn, J. and McCrary, J. (2005). "Incomes in South Africa since the Fall of Apartheid". National Bureau of Economic Research, Working Paper 11384.
- Leibbrandt, M., Poswell, L., Naidoo, P., Welch, M. and Woolard, I. (2006). "Measuring Recent Changes in South Africa Inequality and Poverty using 1996 and 2001 Census Data". In Borhat, H. and Kanbur, R. (eds.) *Poverty and Policy in Post-Apartheid South Africa*. HRSC Press, Pretoria.
- Leite, P.G., McKinley, T. and Osorio, R.G. (2006). "The Post-Apartheid Evolution of Earnings Inequality in South Africa, 1995-2004". United Nations Development Programme, International Poverty Centre, Working Paper 32.
- Majid, N. (2001). "The size of the working poor population in developing countries". Employment Paper 2001/16. International Labour Organization, Geneva.
- Meth, C. and Dias, R. (2004). "Increases in Poverty in South Africa, 1999-2002". *Development Southern Africa*, 21(1): 59-85.
- Posel, D. and Casale, D. (2005). "Who replies in brackets and what are the implications for earnings estimates? An analysis of earnings data from South Africa". Economic Research Southern Africa, Working Paper Number 7, Cape Town.

Potgieter, J.F. (1999). "The Household Subsistence Level in the Major Urban Centres of the Republic of South Africa". Institute for Planning Research Fact Paper No. 107. University of Port Elizabeth, Port Elizabeth.

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.

Rubin, D. B. (1987). *Multiple Imputation for Non-Response in Surveys*. John Wiley and Sons, New York.

Silverman, B. W. (1986). *Density Estimation for Statistics and Data Analysis*. Chapman and Hall, London.

Statistics South Africa (2007). "Consumer price index (CPI)". Statistical Release P0141.1. Statistics South Africa, Pretoria.

van der Berg, S., Burger, R., Burger, R., Louw, M. and Yu, D. (2006). "Trends in poverty and inequality since the political transition". Development Policy Research Unit, Working Paper 06/104.

van der Berg, S., Louw, M. and Yu, D. (2008). "Post-transition poverty trends based on an alternative data source". *South African Journal of Economics*, 76(1): 58 – 76.