

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

Claire Vermaak

University of KwaZulu-Natal

DPRU Annual Conference
27 – 29 October 2008, Muldersdrift

1. Background

- Large literature on levels and trends in post-apartheid poverty using various data sources:
 - 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; Hoogeveen and Özler, 2006)
 - Labour Force Surveys (Meth and Dias, 2004; Leibbrandt *et al*, 2005)
 - more recently, the All Media and Products Surveys (van der Berg *et al*, 2006; van der Berg *et al*, 2008)
- 1995 to 2000: inequality rises but change in poverty more ambiguous. Using hh income or expenditure
- Since 2000: van der Berg *et al* (2008) find that poverty, based on hh per capita income data from the AMPS, has decreased since 2000

- But household surveys usually contain coarsened earnings data:
 - missing values (item non-response);
 - point earnings responses; and
 - interval earnings responses
- Empirical studies on poverty and inequality in SA using LFS or Census data typically ignore missing data, and combine point observations with interval midpoints to create a single earnings variable
- This is problematic:
 - assumes that the data are missing completely at random
 - using interval midpoints ignores the distribution of earnings within intervals
- Ardington *et al* (2006) use impute missing income brackets in Census 2001, but use bracket midpoints for most of their poverty and inequality analysis

2. Research questions

- What is the problem of missing or coarsened data?
- How is multiple imputation implemented to deal with this problem?
- How does the treatment of coarsened data affect estimates of poverty amongst the employed?
- How has poverty amongst the employed changed in SA between 2000 and 2006?

3. Missing and coarsened data

- If the available (observed) data are analysed as if they make up the complete sample:
 - decreased precision results from analysing a smaller dataset
 - inferences may be biased if observed data differ systematically from unobserved data

Types of non-response

- Missing completely at random (MCAR):
 - missingness depends on neither the observed nor the unobserved (missing) data
 - missing data are a simple random sample of the complete dataset
- Missing at random (MAR):
 - missingness depends on the observed data, but not on the unobserved data
 - after conditioning on the observed values, the missing-data structure is ignorable
- Missing not at random (MNAR):
 - missingness depends on both the observed and unobserved data
 - the probability of a value being missing depends on the unobserved value itself, even after conditioning on the observed values

4. Imputation

- Imputation: process of filling in missing data using plausible values.
- Single imputation: replace each missing value with a single predicted value, to create a single complete dataset.
 - Fundamental flaw: fails to take into account that imputed values are more uncertain than observed values. Thus standard errors are likely to be understated, in that they do not reflect this additional uncertainty (Rubin, 1987).
- Multiple imputation: apply a stochastic imputation model to the missing data problem. The model is applied m times, creating m plausible datasets.
 - Produces a distribution of imputed values which reflects the uncertainty involved in the imputation process.

Sequential regression multivariate imputation (SRMI)

- Developed by van Buuren *et al* (1999), extended by Raghunathan *et al* (2001)
- Order the data from the least to the most missing:

$$\underbrace{\mathbf{X}}_{\substack{\text{fully} \\ \text{observed}}}, \underbrace{Y_1, Y_2, \dots, Y_k}_{\substack{\text{some missing} \\ \text{values}}}$$

Cycle 1:

- Regress Y_1 on \mathbf{X} , and impute values for Y_1 using random draws from the appropriate predictive distribution. E.g.
 - normal linear regression model if Y_i is continuous;
 - interval regression model if Y_i contains both missing and interval values (truncated normal distribution for interval values, normal distribution without bounds for missing values)

- Now:

$$\underbrace{\mathbf{X}, Y_1}_{\substack{\text{all values} \\ \text{complete}}} , \underbrace{Y_2, \dots, Y_k}_{\substack{\text{some missing} \\ \text{values}}}$$

- Regress Y_2 on \mathbf{X} and imputed Y_1 , and so on until all Y variables have been imputed, using all previously imputed variables as covariates.
- Cycle 2: Repeat process, updating regression parameters θ with parameters drawn from the now-complete distribution
- Repeat cycles until imputed values and parameters converge to a stable distribution.
 \Rightarrow first imputed dataset

- Entire procedure repeated m times, to produce m imputed complete datasets
- Estimates of interest obtained separately from each of the m imputed datasets are then combined using Rubin's rules:

- Q_i : estimate of interest from the i th imputed dataset, U_i : variance of that estimate.
- Overall combined point estimate is

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

- 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

5. Data sources

- September rounds of the 2000 and 2006 Labour Force Surveys (LFSs):
 - 2000 and 2006 LFS chosen as endpoints due to consistency of survey instrument
 - Period under study includes the 2002 amendment to BCEA, which extended minimum wage determinations number of sectors in which workers traditionally have been vulnerable
- ‘Earnings’ comprises total salary or pay at main job, including overtime, allowances and bonus, before tax or deductions:
 - Likely to understate in-kind benefits
 - Understates total earnings of workers with more than one job
- All earnings figures are in real (2000) prices

Coarsened earnings data in the LFSs

- Proportion of all employed with non-zero working hours reporting earnings as:

	2000	2006
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	0.038 (0.002)	0.067 (0.006)

Notes: Standard errors in brackets

All estimates are weighted to population levels

'Missing' includes the responses 'Don't know', 'Refuse' and 'Unspecified'

6. The imputation model

- Variables included in the model:
 - earnings
 - age and age squared
 - hours worked
 - number of children in the hh
 - dummies for hh head, race, gender, marital status, education, province, metro, occupation, industry, type of business, formal sector
- In particular, natural logarithm of monthly earnings was imputed using interval regression, to deal simultaneously with
 - point observations
 - interval-censored observations
 - missing observations
- SRMI was performed in Stata using the add-on function `ice`:
 - 5 imputed datasets were produced, each using 10 cycles (iterations) of the imputation model

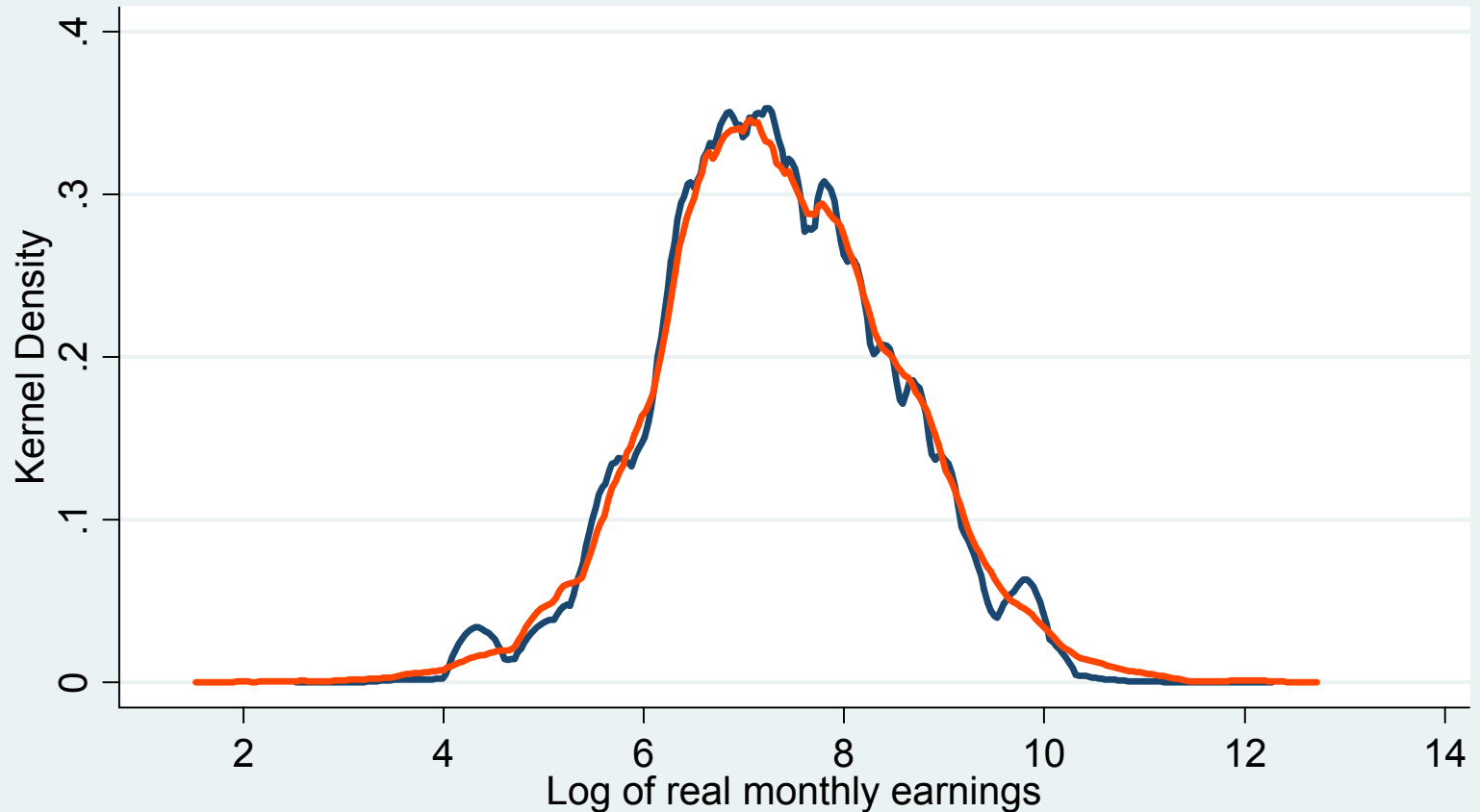
Approaches to the treatment of types of earnings responses

	Treatment of earnings responses			Sample size	Mean of ln(earnings)	Gini	
	Interval	Missing	Zero responses				
Approaches without imputation	A	Midpoints	Omitted	Omitted	24 097	7.308 (0.044)	0.585 (0.009)
	B	Midpoints	Omitted	All included	25 567	6.999 (0.076)	0.602 (0.008)
	C	Midpoints	Omitted	Plausible zeroes included	25 502	7.012 (0.074)	0.602 (0.009)
Approaches using SRMI	D	Imputed	Omitted	Omitted	24 097	7.304 (0.045)	0.590 (0.009)
	E	Imputed	Imputed	Omitted	25 294	7.347 (0.047)	0.634 (0.015)
	F	Imputed	Imputed	All included	26 764	7.056 (0.077)	0.648 (0.014)
	G	Imputed	Imputed	All imputed	26 764	7.293 (0.050)	0.599 (0.010)
	H	Imputed	Imputed	Plausible zeroes included; implausible zeroes imputed	26 764	7.081 (0.074)	0.606 (0.010)

Notes: Standard errors in parentheses

Estimates of mean earnings and Gini coefficients are weighted to population levels

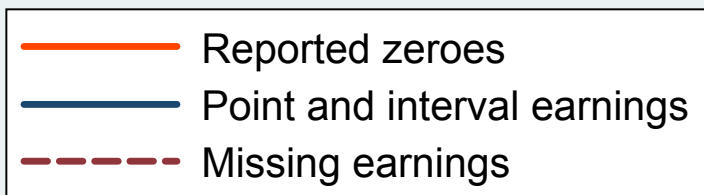
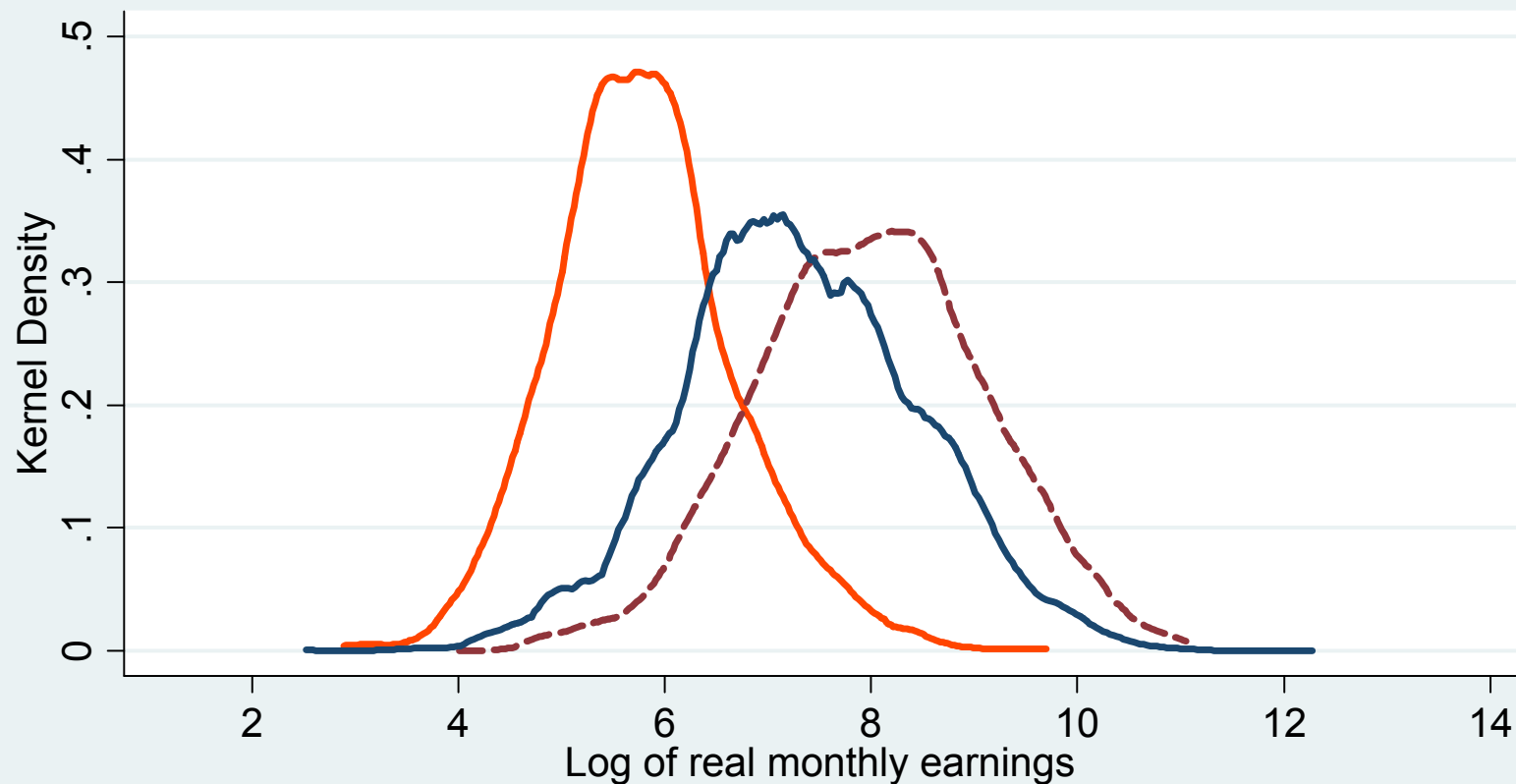
The distribution of earnings, without and with imputation, in 2006



— Without imputation
— With imputed interval and missing data

Note: Estimates are weighted to population levels
Conditional on positive reported earnings

The distribution of imputed zero, missing and interval-censored earnings, in 2006



Note: Estimates are weighted to population levels
Each subsample's density is estimated separately

7. The working poor in SA

- The working poor: those individuals who work, but whose earnings are insufficient to lift them above an individually-defined poverty line. i.e. it is really a study of low-earning workers.
 - Using two poverty lines, in order to:
 - assess the effects of imputation of coarsened data on differently-specified poverty lines
 - assess the extent of changes in poverty at different points in the earnings distribution.
1. R150 per month at real 2000 prices:
 - Corresponds approximately in 2006 to the boundary between the second (R1 – R200) and third (R201 – R500) earnings brackets in the LFSs, in real terms.
 - Close in value to the \$2 per day international poverty line (R159).
 2. R500 per month, at real 2000 prices:
 - Slightly below the midpoint of the fourth earnings bracket (R501 – R1000) in 2006, in real terms.
 - Represents an earnings value approximately 25 percent higher than the household subsistence level per adult equivalent.

Poverty amongst the employed in 2006, without imputation

	Approach		
	A	B	C
	Positive earnings only	Including all zeroes	Including plausible zeroes
Poverty Line 1: R150 per month			
Working poor ('000s)	335 (66)	845 (190)	822 (184)
Headcount ratio	0.029 (0.002)	0.070 (0.007)	0.068 (0.007)
Poverty gap	0.010 (0.001)	0.052 (0.006)	0.050 (0.006)
Poverty Line 2: R500 per month			
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)
Poverty gap	0.065 (0.004)	0.104 (0.009)	0.102 (0.009)
Gini coefficient	0.585 (0.009)	0.602 (0.008)	0.602 (0.009)

Notes: Poverty lines in real 2000 prices
Standard errors in parentheses
All estimates are weighted to population levels

Poverty amongst the employed, by extent of multiple imputation

	Approach		
	A	D	E
	Without imputation: interval midpoints; missing data excl.	Imputation for intervals only; missing data excl.	Imputation for intervals and missing data
Poverty Line 1: R150 per month			
Working poor ('000s)	335 (66)	335 (66)	367 (72)
Headcount ratio	0.029 (0.002)	0.029 (0.003)	0.030 (0.003)
Poverty gap	0.010 (0.001)	0.007 (0.001)	0.010 (0.001)
Poverty Line 2: R500 per month			
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)
Poverty gap	0.065 (0.004)	0.064 (0.004)	0.066 (0.004)
Gini coefficient	0.585 (0.009)	0.590 (0.009)	0.634 (0.015)

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

Poverty amongst the employed, by method of imputation of reported zero earnings

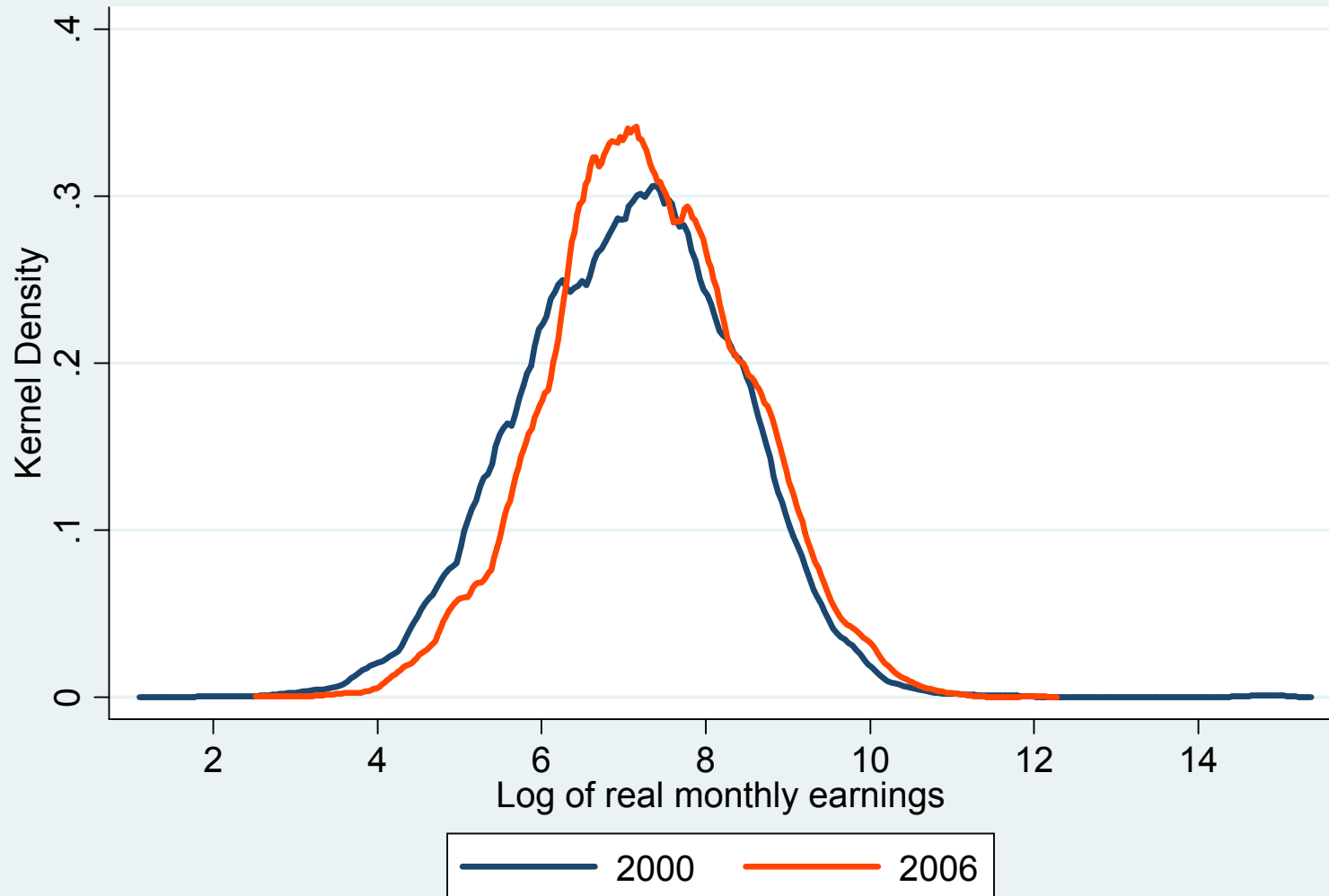
	Approach		
	F	G	H
	All zeroes included	All zeroes imputed	Implausible zeroes imputed
Poverty Line 1: R150 per month			
Working poor ('000s)	877 (195)	425 (93)	815 (182)
Headcount ratio	0.068 (0.007)	0.033 (0.003)	0.063 (0.007)
Poverty gap	0.049 (0.006)	0.008 (0.001)	0.043 (0.005)
Poverty Line 2: R500 per month			
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)
Poverty gap	0.103 (0.009)	0.072 (0.005)	0.096 (0.008)
Gini coefficient	0.648 (0.014)	0.599 (0.010)	0.606 (0.010)

Notes: Poverty lines in real 2000 prices

Standard errors in parentheses

All estimates are weighted to population levels

The distribution of real log monthly earnings, 2000 and 2006



Note: Estimates are weighted to population levels

8. Poverty levels and trends, 2000 and 2006

- Substantial number of workers earn less than poverty line
- Significant decline in poverty amongst the employed over time
- Greater improvement at bottom of earnings distribution

Poverty Line 1: R150 per month	2000	2006	Change (%)
Working poor ('000s)	1 387 (65)	815 (182)	-41.2
Headcount ratio	0.114 (0.005)	0.063 (0.007)	-44.4
Poverty gap	0.097 (0.005)	0.043 (0.005)	-55.1
<hr/>			
Poverty Line 2: R500 per month			
Working poor ('000s)	3 304 (89)	2 415 (470)	-26.9
Headcount ratio	0.271 (0.007)	0.187 (0.014)	-30.8
Poverty gap	0.167 (0.005)	0.096 (0.008)	-42.7

Notes: Standard errors in parentheses
Missing earnings and implausible zeroes imputed

- Analysis presented here is merely suggestive:
 - What are the correlates of low-earning work?
 - 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?
- Specific focus on earnings is useful (enables an analysis of the effects of labour market trends and policies on poverty separate from the effects of the extension of the social welfare system)
- But also need to link low-earning workers with other sources of income in their households, in order to assess overall poverty outcomes

9. Summary

- Imputing coarsened data using SRMI does not significantly affect estimates of poverty rates amongst the employed:
 - imputed values for missing earnings observations mostly fall above the poverty line
 - imputation of interval responses affects estimates of poverty rates only to the extent that the poverty line bisects an interval
- How zero-earning workers are treated makes a far greater difference to estimates of poverty than does the treatment of missing and interval-reported data
- Multiple imputation provides an attractive method of dealing with coarsened survey data
- But it is costly in terms of time and computing resources
- Estimates of poverty amongst the employed are not significantly different when implementing SRMI than when using more traditional methods
- Do the benefits of multiple imputation outweigh its costs, and should this methodology become standard practice amongst poverty researchers?