DataFirst Technical Papers

Nonparametric estimation when income is reported in bands and at points

by

Martin Wittenberg

Technical Paper Series
Number 8

About the Author(s) and Acknowledgments

I would like to thank Anthea Heap, Reza Daniels and participants at a SALDRU seminar. Economic Research Southern Africa (ERSA) provided financial support as well as a useful set of comments. Contact details: Martin.Wittenberg@uct.ac.za

Recommended citation

Wittenberg, M. (2008). Nonparametric estimation when income is reported in bands and at points. A DataFirst Technical Paper Number 8. 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

Nonparametric estimation when income is reported in bands and at points*

Martin Wittenberg
School of Economics and SALDRU
University of Cape Town

September 2008

Abstract

We show how to estimate kernel density functions of distributions in which some of the responses are provided in brackets, by inverse probability weighting. We consider two cases, one where the data are CAR and where the data are not CAR. We show how the selection probabilities can be estimated by means of the EM algorithm without specifying a parametric distribution function for the variable. A Monte Carlo experiment shows that this procedure estimates the selection parameters fairly precisely. We apply these techniques to earnings data from South Africa's first post-apartheid nationally representative survey, the 1994 October Household Survey.

Keywords: coarsening, bracket responses, EM algorithm, inverse probability weighting
JEL codes: C13, C14

*I would like to thank Anthea Heap, Reza Daniels and participants at a SALDRU seminar. Economic Research Southern Africa (ERSA) provided financial support as well as a useful set of comments. Contact details: Martin.Wittenberg@uct.ac.za

1 Introduction

In many surveys it is customary to provide individuals with the option of providing their income information in bands if they do not wish to volunteer a point estimate. This raises the question as to how to combine the information in the categorical variable with that given as point values. A common practice is to impute the categorical information by placing the individuals at some point in the band, e.g. at the midpoint. In the case of the "open" interval some other rule is used, e.g. some fixed multiple of the lower limit of that band is used.

Putting every individual with missing point information at the same point in the distribution creates difficulties if the estimation is concerned with anything other than means. In particular, if one is concerned with kernel density estimation these spikes are very difficult to smooth over. One approach is to place individuals randomly within the category, using some pre-specified distribution, e.g. uniform or log-normal (e.g. Woolard and Woolard 2006). This creates additional difficulties, since one is now adding in a random error to the imputed value. In general the two errors will not offset each other. One way around this is to impute the value several times and perform the analyses on multiple data sets. Another advantage of multiple imputation is that it explicitly recognises the fact that the imputed values are not data of the same quality as the point values. The uncertainty associated with the imputation process is therefore explicitly recognised.

In this paper we suggest that under some circumstances it may be even better to use a reweighting strategy. The information from the categorical variable is used to calculate the appropriate weights to apply to the point values. We will show that this strategy is particularly attractive if the spirit of the enquiry is nonparametric, i.e. if one wanted to estimate the distribution of the variable without imposing a particular parametric form. This strategy will outperform either simple or multiple imputation whenever the distribution of the variable in question is markedly different from the distributional assumptions implicit in the imputation strategy. We will show this in some detail below.

Reweighting as a strategy is not a novel concept. It is used frequently when dealing with unit nonresponse. However, it is used less frequently when the issue is item nonresponse, indeed in parts of the literature the conventional wisdom is that "weighting adjustments are made for unit nonresponse and imputation is employed for item nonresponse" (Duncan and Kalton 1987, p.110). A bracket response can be seen as analogous to item nonresponse. Our discussion can therefore also be seen as broadening the discussion about reweighting adjustments.

The structure of the paper is as follows: we begin with a review of the literature in section 2 and then present our model in section 3. We initially allow the data to be "coarsened at random" and show that reweighting the observations will allow consistent estimators of all the moments of the distribution as well as of the distribution function itself. In Section 4 we consider a more realistic model, where the data are no longer CAR. Under these circumstances the simple reweighting strategy will lead to biased and inconsistent results. We assess the impact of ignoring this problem in Section 4.1, by means of a Monte Carlo experiment. We then demonstrate, in Section 4.2 that we can, in fact, improve on the simple reweighting strategy if we estimate the underlying coarsening mechanims parametrically. We do so by means of the EM algorithm. We again provide some Monte Carlo evidence that this approach works well. In the final section we apply these approaches to an empirical example, the analysis of the distribution of earnings in South Africa's first nationally representative household survey of the post-apartheid era, the 1994 October Household Survey. We show that the estimated mean increases by 23% when we reweight the data assuming that the data has been coarsened at random. Estimating the coarsening mechanism by the EM algorithm leads to an additional increase in mean income by anywhere from 2% to 15%. We show, however, that this parametric reweighting can distort the resulting distribution, so that caution in using this method is called for.

2 Weighting as a response to missing information

The idea of reweighting the observed data to account for missing information is not new. Indeed it is done routinely by statistical agencies in order to deal with unit nonresponse (for discussions see Elliot 1991, Holt and Elliot 1991, Kalton and Flores-Cervantes 2003). Much of this literature, however, does not consider the problem of item nonresponse. The literature dealing with "missing data" and data imputations (e.g. Schafer 1997) on the other hand tends to ignore the possibility of weighting adjustments. However, as Little (1988) notes, there are important connections between the two: for instance a "hotdeck" imputation can be achieved by increasing the weight on an existing observation to cover the imputed case; and multiple imputations could be achieved by the creation of multiple weights.

More recently Wooldridge (2007) showed that "inverse probability weighting" could lead to consistent estimation of parameters in a range of contexts largely (though not exclusively) where the data have been

coarsened at random. The approach that we outline below is similar in spirit although it differs in two important respects from this work. In the first instance our objective is not to estimate a set of structural parameters. Secondly we show that it may be possible to estimate the coarsening probabilities when the data are non-CAR, without specifying a parametric family of distributions for the coarsened variable.

Central to any reweighting strategy is the calculation of "response propensities". These are the probabilities of responding to a particular survey or to a particular question on that survey. This propensity can be modelled, for instance, by means of a logit or probit model. The appropriate weights are then proportional to the inverse of these propensities. Little (1988) points out the connection between this approach and the theory of propensity scores of Rosenbaum and Rubin (1983). He points out also several problems with the reweighting approach including the fact that the range of the weights thus calculated can lead to an increase in the variance of the estimates. Indeed we will show this in some of our simulations below. The problem attendant on extreme weights is recognised more generally in the literature. Kalton and Flores-Cervantes (2003, p.90) note that it is common practice to "trim" such weights. Little (1988, p.293) comments that:

> This method may reduce mean squared error, but it is ad hoc with little or no theory to substantiate the choice of cut-off.

Additional problems noted by Little (1988) are that correct inference with weighted data can be problematic and in the case of item nonresponse the complexity of the patterns of missing data cause difficulty. In the case we are analysing below, however, we are essentially dealing with a univariate type of analysis, so that we do not have to contend with the problem of data missing in different parts in different variables. In order to make this discussion more concrete we now outline our basic model.

3 The basic model

Assume that the random variable X has distribution function f, but that we do not fully observe the draws from this distribution, i.e. we observe either x or the interval into which x falls, i.e. $x \in Z_j$, where the set of disjoint intervals Z_j, $j \in \{1, 2, \ldots, m\}$ cover the range of X[1]. Following Heitjan and Rubin (1991) we represent the observed outcome as $Y = Y(X)$, where the sample space of Y is the power set of the sample space of X. Let the random variable G define the event that individual i reports an income amount, i.e. this is a dummy variable with realisations $g = 1$ if the individual provides a point value and $g = 0$ if not. If $g = 1$, then $y = \{x\}$ while if $g = 0$, $y = z(x)$, the interval into which x falls.

Let the density of the observed x values be $f^{obs}(x)$. By definition this is

$$\begin{aligned} f^{obs}(x) &= f(x|g=1) \\ &= \frac{f(x)\Pr(g=1|x)}{\Pr(g=1)} \end{aligned}$$

We make the key assumption that we can invert this relationship.

Assumption 1 *We can write the density of the partially observed X variable as*

$$f(x) = \frac{\Pr(g=1)}{\Pr(g=1|x)} f^{obs}(x) \tag{1}$$

*i.e. we assume that there are no x values in the range of X where $\Pr(g=1|x) = 0$. In particular this means that the sample space of Y conditional on $g = 1$, i.e. the sample space of the **observed** x values, covers the full range of X. This means that there are no parts of the range of X which we will only observe as brackets. We are therefore explicitly ruling out certain forms of censoring, e.g. "topcoding" of large values.*

Equation 1 can be written as

$$f(x) = w(x) f^{obs}(x)$$

where $w(x)$ is a weighting function. The weight is inversely proportional to the probability of x being observed, so this is a straightforward case of "inverse probability weighting". If the $w(x)$ function depends only on the observed information y, the variable X is "coarsened at random" (CAR) and we can consistently estimate all moments of the distribution f by reweighting the observed point values.

[1] We are ignoring cases where information is completely missing.

Bracket	$Pr(g=1)$
$1-100$	0.62
$100-200$	0.70
$200-500$	0.67
$500-1000$	0.45
$1000-2000$	0.36
$2000-4000$	0.32
$4000-8000$	0.30
$8000-16000$	0.29
16000+	0.25

Table 1: Brackets and probabilities used in the Monte Carlo experiments.

Example 2 *Assume that* $\Pr(g=1|x) = \Pr(g=1|x \in Z_j)$*, i.e. the probability of providing a point value depends only on the bracket* Z_j*, but not on the precise point within the bracket, then* $w(x) = w(y)$ *and the variable* X *is CAR. If we have a simple random sample of size* n *from the distribution of* X*, then under standard conditions a consistent estimator of* $E(X)$ *will be given by*

$$\widehat{x}_w = \sum_{ij} x_{ij}^{obs} \frac{1}{n\widehat{p}_{1,j}} \tag{2}$$

where we are taking the sum over point values, x_{ij}^{obs} *is the* i*-th observation in category* Z_j *and* $\widehat{p}_{1,j}$ *is the proportion of point responses in category* Z_j*. The formula given in equation 2 is equivalent to the Horvitz-Thompson estimator for sample surveys (Horvitz and Thompson 1952). Indeed, we could think of the process of providing point estimates as a "sampling" process from the potential* x *values available in our finite data set. Of course, as Little (1988) notes, the same point estimate of* $\widehat{x}_w$ *will be achieved by imputing all the "missing" responses to the category means of the bracket into which they fall.*

Such mean imputations will not estimate higher moments consistently. Furthermore they will produce spikes which distort the standard kernel regression estimators. Furthermore these spikes do not disappear with additional information, since the process leading to category only information is intrinsic to the data sampling process. Even in very large samples there will be a non-zero fraction of the observations which will arrive in categorical form only.

In the case where the X variable is coarsened at random, we can, however, produce consistent estimates of the densities using weighted forms of kernel regression on the point values. This is evident from equation 1, where the weights will be $w(y)$, which can be estimated from the data. In the case of the example given above, the weights will be inversely proportional to the estimated probabilities $\widehat{p}_{1,j}$ of providing point responses.

3.1 Monte Carlo experiments

In order to explore the performance of this reweighting strategy we run a series of Monte Carlo experiments. In all cases we draw samples of size 30,000 from a specified distribution (lognormal or a mixture of lognormal distributions). We coarsen these data, initially according to constant probabilities within each "bracket". For the purposes of these experiments we define the brackets as given in Table 1, which also gives the associated probability of providing a point estimate. These categories and the associated reporting probabilities are loosely based on the 1994 October Household Survey. The pattern (with the exception of the lowest bracket) shows a clear decrease in the probability of giving point estimates as income increases.

We report the performance of four different estimators in estimating four moments, the mean, variance, μ_3 and μ_4, where the latter two are the third and fourth moments about the mean. We also report the root mean square error, to facilitate comparison between the estimators. The estimators are:

1. The "reweighting" estimator, which uses the reweighting procedure outlined in equation 2 above.
2. The estimator obtained when the bracket responses are simply replaced by the mean of the observed values within the bracket.
3. An estimator which randomly imputes the missing values within each bracket using the lognormal distribution. The parameters of the distribution are estimated using the reweighted mean and reweighted variance of the lognormal distribution, calculated according to the first estimator. The imputation procedure is repeated ten times and the results computed are the means of these ten imputation runs.

True distribution:	Lognormal				Mixture			
	μ	σ^2	μ_3	μ_4	μ	σ^2	μ_3	μ_4
Parameter value	1636.0	4.60E+06	6.10E+10	2.41E+15	8571.9	6.49E+08	1.75E+14	1.34E+20
Reweighted	1636.0	4.60E+06	6.04E+10	2.19E+15	8569.6	6.48E+08	1.74E+14	1.24E+20
	(13.6)	(4.61E+05)	(4.57E+10)	(6.54E+15)	(242.0)	(1.27E+08)	(2.11E+14)	(5.15E+20)
RMSE	13.6	4.61E+05	4.57E+10	6.54E+15	242.0	1.27E+08	2.11E+14	5.15E+20
Imputed mean values	1636.0	4.31E+06	4.47E+10	1.05E+15	8569.6	3.75E+08	5.12E+13	3.14E+19
	(13.6)	(3.10E+05)	(1.62E+10)	(1.71E+15)	(242.0)	(4.36E+07)	(5.38E+13)	(1.31E+20)
RMSE	13.6	4.25E+05	2.30E+10	2.18E+15	242.0	2.78E+08	1.35E+14	1.66E+20
Multiple imputations - normal	1635.9	4.59E+06	6.03E+10	2.24E+15	7844.9	7.02E+08	6.40E+14	4.29E+21
	(12.4)	(2.08E+05)	(1.46E+10)	(2.31E+15)	(154.6)	(1.28E+08)	(2.07E+15)	(4.57E+22)
RMSE	12.4	2.08E+05	1.46E+10	2.32E+15	743.2	1.38E+08	2.12E+15	4.59E+22
Multiple imputations - uniform	1720.8	5.49E+06	2.77E+11	4.22E+17	19432.3	3.93E+14	8.03E+25	1.65E+35
	(13.8)	(1.51E+06)	(3.49E+12)	(1.06E+19)	(36275.8)	(7.71E+15)	(1.75E+27)	(4.41E+36)
RMSE	86.0	1.75E+06	3.50E+12	1.06E+19	37866.6	7.72E+15	1.75E+27	4.41E+36

Notes:
1000 replications of estimates from a sample of size 30,000. The lognormal distribution is LN(6.9,1).
The mixture is $z * x_1 + (1 - z) * x_2$ where x_1 is LN(6.9,1), x_2 is LN(10,1) and z a Bernoulli r.v. with $p = 0.8$
$\mu_3 = E\left[(X - \mu)^3\right]$ and $\mu_4 = E\left[(X - \mu)^4\right]$
Standard deviations of the Monte Carlo estimates are given in parentheses

Table 2: Monte Carlo experiments: Estimation of Moments

4. The final estimator randomly imputes the missing values within each bracket using a uniform distribution. Since we are again repeating this procedure ten times, this is really equivalent to putting individuals at the midpoint of each bracket, when we consider estimating the mean. We cannot use this procedure for the open bracket, so there we impute the missing values according to a Pareto distribution. This amounts to putting individuals (on average) at a fixed multiple of the bracket boundary. The parameter of the Pareto distribution is estimated from the observed values in the top category by maximum likelihood.

In Table 2 we show the results when these estimators are applied to two different types of distributions. In the left hand panel we consider draws from a lognormal distribution, specifically a $LN\,(6.9, 1)$ distribution. The right hand panel considers a density which is a random mixture of two lognormal distributions, one $LN\,(6.9, 1)$ and the other $LN\,(10, 1)$. The probability of the former being observed is 0.8. This density is arguably closer to South African income distributions which exhibit some degree of bimodality.

Considering the draws from the lognormal distribution $LN\,(6.9, 1)$ first, i.e. the left panel of Table 2, we observe that the reweighting procedure is highly successful in estimating the mean and centred moments of the lognormal distribution. As expected, the estimated mean of the second estimator is identical to that obtained by the reweighting procedure and the higher moments show considerably more bias than the reweighting estimator. This is offset by the smaller variability of the estimates, resulting in an overall lower root mean square error. The "multiple imputation" estimator which imputes values according to the lognormal distribution performs best according to the root mean square error criterion. The final estimator which imputes randomly within brackets and according to a Pareto distribution in the upper tail performs badly. How this estimator might perform with more thick-tailed distributions is an open question.

In the second experiment (reported in the right most columns of Table 2) where we estimate moments of the mixture of lognormal distributions, the "reweighting" estimator outperforms the three estimators that use imputations.

We gain additional insight into this case by a third experiment, in which we used the first, third and fourth approaches to estimate kernel densities of the log of income of the same mixed distribution considered above. We did not use the second approach, since the conversion of brackets to point estimates creates spikes in the distribution which create problems for the density estimators. We estimated the density at a regularly spaced grid of sixty points. The means of 1000 replications of these estimates are graphically depicted in the top panel of Figure 1. The first point which stands out is that the density of the observed values is clearly a very poor approximation to the true density. However, it is also evident that on average the "reweighting" procedure tracks the true density remarkably well; indeed, much more successfully than the estimates which impute point values to the bracket responses. The reason for this is clear when we take note of the placement of the top bracket boundary (indicated by the vertical line). The imputation procedures tend to put the imputed values in the top bracket too close to that boundary. The reason for this is that the shape of the parametric distribution that these procedures try to impose on the data is simply wrong for that bracket.

The superior performance of the (log)normal imputations in the left panel of Table 2 is due to the fact that the assumption of normality was correct. The assumption therefore added useful information about where

the missing observations were located that the reweighting procedure did not have access to. When that assumption was mistaken, however, the performance was significantly worse, as in the right panel of Table 2.

The reason for the "bump" immediately to the left of the vertical line in the top panel of Figure 1 is that the kernel density estimators smooth over the edges of that bracket boundary. In the bottom panel of Figure 1 we show that the artificial mode created by the imputation procedure is not a mere statistical accident. We connect up the 2nd percentiles of the Monte Carlo estimates and the 98th percentiles respectively. These bounds provide non-parametric 96% confidence intervals for the density estimates at each of the 60 points of the estimation grid. As the figure shows, the true density is always within the bounds given by the reweighting estimator. This is not the case, however, for the multiple imputations using the normal distribution. This indicates that the latter estimator is biased. Since the proportion of bracket values does not go to zero as the sample size increases it is obvious also that the estimator will be inconsistent.

4 Estimation when the data are not coarsened at random

Let us now consider a situation in which the coarsening mechanism does **not** operate at random. In particular, let us assume that the probability of giving a point response is a function of income itself, more particularly that it is a probit model with parameter vector γ:

$$\Pr(g = 1|x, \gamma) = \Phi(x\gamma)$$

Consequently the joint distribution of (g, x), is

$$\phi(g, x|\gamma) = [\Phi(x\gamma)]^g [1 - \Phi(x\gamma)]^{1-g} f(x)$$

The distribution of the observed data y will be

$$P(y|\gamma) = \int_{g,x\in y} \phi(g, x|\gamma)\, dx$$

We have assumed throughout that the outcome g is observed, so

$$P(y|\gamma) = \begin{cases} \Phi(x\gamma) f(x) & \text{if} \quad y = \{x\} \\ \int_{x\in Z_j} [1 - \Phi(x\gamma)] f(x)\, dx & \text{if} \quad y = Z_j \end{cases}$$

The log-likelihood based on the observed data is

$$L(\gamma|y) = \log P(y|\gamma) \tag{3}$$

We will show below that we can maximise this by means of the EM algorithm under certain conditions.

Of course if we were able to observe the data completely, it would be relatively trivial to estimate γ. Assuming independent random sampling, the complete data likelihood function will be

$$L(\gamma|x, g) = \prod_{i=1}^{n} [\Phi(x_i\gamma)]^{g_i} [1 - \Phi(x_i\gamma)]^{1-g_i}$$

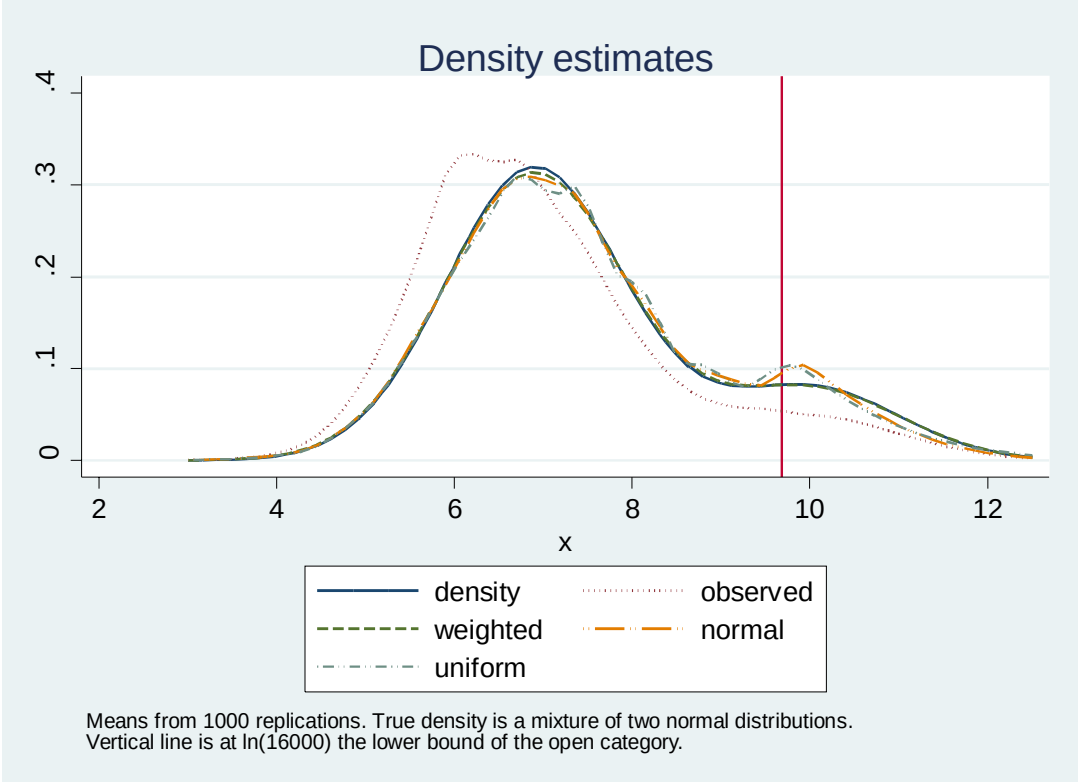
where we have used the conditional distribution $h(g|x, \gamma)$ as the basis for the likelihood. The corresponding log-likelihood is

$$\ln L(\gamma|x, g) = \sum_{i=1}^{n} g_i \ln [\Phi(x_i\gamma)] + \sum_{i=1}^{n} (1 - g_i) \ln [1 - \Phi(x_i\gamma)] \tag{4}$$

This could be estimated in the standard way, **if** we had point values for all x. Since, however, x_i is missing (recorded only in brackets) if $g_i = 0$, we have no non-zero contributions to the second term.

4.1 Ignoring the non-CAR nature of the data

An initial response might be to simply ignore the fact that the data are no longer coarsened at random and to proceed with the reweighting procedure outlined in the previous section. It is clear that this will introduce an element of bias, since the probability of reporting information in brackets is no longer constant within categories. Giving the same weight to all points within a bracket will tend to underweight the high incomes



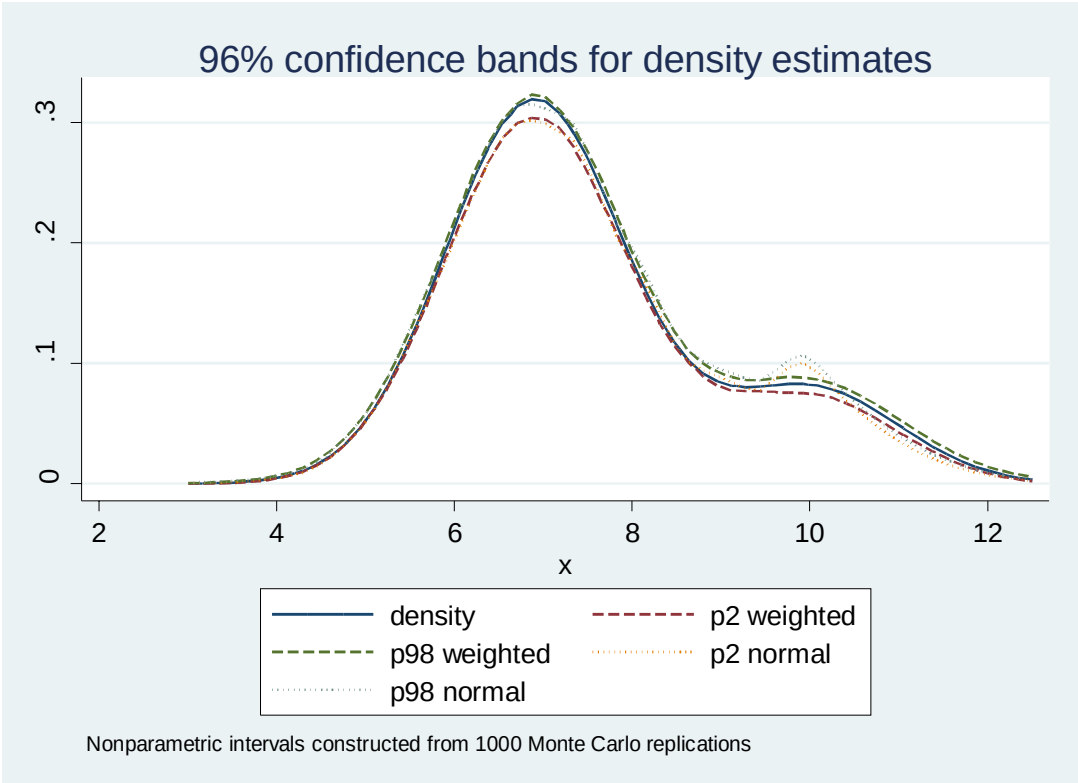


Figure 1: Estimates of the density of the true distribution – results from a Monte Carlo experiment. Top panel: the "reweighted" estimator provides a reasonably unbiased estimate of the true density. The imputation procedures create an artifical second mode in the distribution. Bottom panel: The true density lies between the 2nd and 98th percentile of the estimates from the "reweighted" estimator, but the imputations using the normal distribution do not capture the true distribution in the top tail.

True distribution:	Lognormal				Mixture			
	μ	σ^2	μ_3	μ_4	μ	σ^2	μ_3	μ_4
Parameter value	1636.0	4.60E+06	6.10E+10	2.41E+15	8571.9	6.49E+08	1.75E+14	1.34E+20
Reweighted	1625.9	4.51E+06	5.79E+10	2.31E+15	7847.8	4.83E+08	1.04E+14	5.79E+19
	(13.0)	(4.67E+05)	(7.05E+10)	(1.69E+16)	(209.8)	(9.15E+07)	(1.20E+14)	(2.55E+20
RMSE	16.5	4.76E+05	7.06E+10	1.69E+16	753.9	1.90E+08	1.40E+14	2.66E+20
Imputed mean values	1625.9	4.24E+06	4.34E+10	1.09E+15	7847.8	2.85E+08	2.95E+13	1.35E+19
	(13.0)	(3.06E+05)	(2.35E+10)	(5.02E+15)	(209.8)	(3.06E+07)	(2.76E+13)	(5.72E+19)
RMSE	16.5	4.69E+05	2.94E+10	5.19E+15	753.9	3.66E+08	1.48E+14	1.33E+20
Multiple imputations - normal	1632.2	4.58E+06	6.05E+10	2.41E+15	7527.3	6.06E+08	5.10E+14	2.74E+21
	(12.0)	(2.12E+05)	(2.28E+10)	(5.68E+15)	(144.7)	(1.05E+08)	(1.13E+15)	(1.62E+22)
RMSE	12.6	2.13E+05	2.28E+10	5.68E+15	1054.6	1.14E+08	1.18E+15	1.64E+22
Multiple imputations - uniform	1712.3	5.39E+06	3.98E+11	1.05E+18	13547.9	1.58E+14	2.81E+25	2.84E+34
	(13.3)	(2.00E+06)	(6.92E+12)	(2.67E+19)	(22923.6)	(4.30E+15)	(8.66E+26)	(5.65E+35)
RMSE	77.4	2.15E+06	6.93E+12	2.67E+19	23457.5	4.30E+15	8.66E+26	5.66E+35

Notes:
1000 replications of estimates from a sample of size 30,000. $\Pr(g=1|x,\gamma)=\Phi(0.85-0.15x)$
The lognormal distribution is LN(6.9,1).
The mixture is $z*x_1+(1-z)*x_2$ where x_1 is LN(6.9,1), x_2 is LN(10,1) and z a Bernoulli r.v. with $p=0.8$
$\mu_3=E\left[(X-\mu)^3\right]$ and $\mu_4=E\left[(X-\mu)^4\right]$
Standard deviations of the Monte Carlo estimates are given in parentheses

Table 3: Monte Carlo experiments: Estimation of Moments ignoring that the data are nonCAR

(assuming that $\gamma<0$) and overweight observations near the lower boundary of the bracket. Nevertheless if the brackets are reasonably small one might assume that the bias introduced in this way may be tolerable.

We explore this by a Monte Carlo experiment, again drawing samples of size 30,000 from a lognormal distribution and then again from a mixed distribution with the same brackets as before. The parameters of the probit model were set at $\gamma_1=.85$ and $\gamma_2=-0.15$. With these parameters the probability of providing point estimates decreases from an average of around 0.65 in the lowest bracket to 0.25 in the top bracket, which is similar to the probabilities reported in Table 1.

The results of the experiment, shown in Table 3, suggest that the performance of the estimators depends on how much weight there is in the upper tail of the distribution. The left-hand panel of Table 3 should be compared to the left-hand panel of Table 2. It is evident that the reweighting estimator tends to underestimate the higher moments of the data, but the impact of the misspecification is not as dramatic as one might have supposed. This is in strong contrast to the right-hand panel of Table 3, where we now see a noticeable bias in the estimation of the mean when using the reweighting estimator, so that the Root Mean Square Error increases threefold (from 242 to 754). The estimators which impute values for the missing point values also perform badly on the mixed distribution. Indeed, they do worse. The fact that the data may not be CAR is therefore no argument for doing the simple imputation procedures discussed above.

It turns out, however, that we can improve on the simple reweighting estimator by parametrically estimating the probability of coarsening.

4.2 Estimating the coarsening mechanism parametrically

We noted above that the log-likelihood (given in equation 4) could not be estimated in the standard way, since the x values were all missing for the case where $g_i=0$. Nevertheless we can estimate γ in that equation by means of the EM algorithm provided that the distribution of the random variable X does not itself depend on γ.

Assumption 3 *The distribution function of X, i.e. $f(x)$ is independent of γ.*

We will follow the exposition in Dempster, Laird and Rubin (1977). Let

$$k(g,x|y,\gamma)=\frac{\phi(g,x|\gamma)}{P(y|\gamma)} \tag{5}$$

It then follows that

$$L(\gamma)=\log\phi(g,x|\gamma)-\log k(g,x|y,\gamma)$$

Define

$$\begin{aligned} Q(\gamma|\gamma^p) &= E[\log\phi(g,x|\gamma)|y,\gamma^p] \\ &= \begin{cases} \log\Phi(x\gamma)+\log(f(x)) & \text{if} \quad y=\{x\} \\ E\{\log[1-\Phi(x\gamma)]|y,\gamma^p\}+E[\log(f(x))|y,\gamma^p] & \text{if} \quad y=Z_j \end{cases} \end{aligned} \tag{6}$$

and let

$$H(\gamma|\gamma^p) = E[\log k(g,x|y,\gamma)|y,\gamma^p] \tag{7}$$

The generalised EM algorithm (Dempster et al. 1977, p.7) now consists of the following steps:

1. The Expectation or E-step:

   Let γ^p be a preliminary estimate of γ. Then calculate $Q(\gamma|\gamma^p)$ for all $\gamma \in \Omega$ where Ω is the parameter space

2. The Maximisation or M-step:

   Find γ^{p+1} by maximising $Q(\gamma|\gamma^p)$ as given in equation 6, i.e.

$$y^{p+1} = \arg\max_{\gamma} Q(\gamma|\gamma^p)$$

Beginning with an arbitrary starting point $y^{(0)}$ we generate the sequence of estimates $\left\{y^{(0)}, y^{(1)}, y^{(2)}, \ldots, y^{(p)}\right\}$ until the series has converged.

As it stands, this is not a well-defined algorithm, since we have not specified what to do about the $\log(f(x))$ terms in $Q(\cdot|\cdot)$. Furthermore we need to be able to calculate $E\{\log[1-\Phi(x\gamma)]|y,\gamma^p\}$. However by our assumption $\log(f(x))$ is independent of γ, i.e. it is effectively a constant and will drop out of the maximisation process. In practice we can maximise

$$Q^*(\gamma|\gamma^p) = \begin{cases} \log\Phi(x\gamma) & \text{if} \quad y=\{x\} \\ E\{\log[1-\Phi(x\gamma)]|y,\gamma^p\} & \text{if} \quad y=Z_j \end{cases}$$

and this will maximise $Q(\gamma|\gamma^p)$.

In order to calculate $E\{\log[1-\Phi(x\gamma)]|y,\gamma^p\}$ we make use of assumption 1. Writing

$$f(x) = w(\gamma) f^{obs}(x)$$

we take the expectation in equation 6 (and indeed in 7) with respect to the observed distribution of x in interval Z_j, i.e.

$$E\{\log[1-\Phi(x\gamma)]|y,\gamma^p\} = \int_{x\in y} \log[1-\Phi(x\gamma)]\, w(\gamma^p) f^{obs}(x)\, dx \tag{8}$$

This process is guaranteed to lead to increases in $L(\gamma)$ for the same reasons outlined in the original Dempster et al. (1977) article: $H(\gamma|\gamma^p) \leq H(\gamma^p|\gamma^p)$ except if $k(g,x|y,\gamma) = k(g,x|y,\gamma^p)$ almost everywhere; and $Q\left(\gamma^{p+1}|\gamma^p\right) \geq Q(\gamma|\gamma^p)$, so at each step of the process $L(\gamma|y)$ increases.

We make one more adjustment to the procedure. In practice we calculate the right hand side of equation 8 not with respect to the true density f^{obs}, but a consistent estimate of it, the sample distribution function $\widehat{f}^{obs}$. This does not affect any of the convergence properties of the algorithm. It does mean, however, that we are implicitly looking for a maximum of

$$L(\gamma|y) = \log \int_{x\in y} [\Phi(x\gamma)]^g [1-\Phi(x\gamma)]^{1-g} w(\gamma) \widehat{f}^{obs}(x)\, dx$$

We show in the Monte Carlo experiment below that this procedure seems to produce consistent results.

Once we have an estimate of γ, we can reweight the observed density $f^{obs}(x_i)$ to generate an estimate of $f(x_i)$ along the lines shown in equation 1. The parametrically estimated weights therefore enable us to estimate the density (or any of the moments) nonparametrically.

	Estimate	Std.Error		
$\widehat{\gamma}_1$	0.8388041	(0.0366526)		
$\widehat{\gamma}_2$	−0.1485911	(0.0048508)		
	μ	σ^2	μ_3	μ_4
true value	8571.9	6.49E+08	1.75E+14	1.34E+20
Weighted Estimate	8513.2	6.54E+08	2.65E+14	7.01E+20
	(322.2)	(5.64E+08)	(3.21E+15)	(1.90E+22)
RMSE	327.5	5.64E+08	3.21E+15	1.90E+22
Notes:				
The distribution is $z * x_1 + (1 - z) * x_2$ where x_1 is LN(6.9,1),				
x_2 is LN(10,1) and z a Bernoulli r.v. with $p = 0.8$				
True parameters: $\gamma_1 = 0.85$, $\gamma_2 = -0.15$				

Table 4: Monte Carlo experiment: Estimation of γ by EM algorithm and then reweighting

4.3 Monte Carlo results

In order to explore how successful this procedure is we ran another Monte Carlo experiment in which we again drew 1000 samples of size 30,000 from the mixture of lognormal distributions considered earlier. On these samples we first estimated the parameters of the coarsening mechanism and then used the resulting estimate to reweight the observations according to equation 1. The results are given in Table 4.

The top panel of this table shows that the EM algorithm succeeded remarkably well in estimating the true parameters of the model. The mean values of the point estimates are close to the true values and the Monte Carlo standard errors are fairly small, suggesting that most samples produced estimates close to the true values. The bottom panel shows what happens to our estimates when we reweight using the appropriate estimates of the probability of coarsening. It is evident that the mean of the distribution is estimated much more precisely (compare to the right hand panel of Table 3). The RMSE has more than halved from 754 to 328. Our simulations suggest that the bias of higher moments is reduced, but this is offset by greater variability of the estimates[2]. Indeed the RMSE of our estimate of σ^2 has doubled. The reason for this is due to the fact that the estimation of higher moments will be more sensitive to extreme weights. Reweighting each point estimate proportional to the inverse of the probability of it being observed will potentially lead to some enormous weights. As it turns out, the reason for the higher variance of our Monte Carlo estimates of the higher moments is due entirely to one simulated data set in which the maximum value of x (over R5 million) also happened to be reported as a point value. The next highest reported value, for comparison, was 374,000. The probability of this extreme value being reported as point estimate according to our probit model was 6%, so this outlier was accorded a large weight, leading to a very high estimate of σ^2, μ_3 and μ_4 on that sample. The estimator that applied constant weights within brackets, by contrast, did not weight this particular observation to the same extent.

In the empirical example that we consider below we do not observe such extreme values. Nevertheless the lesson from the extreme Monte Carlo sample is that one should be cautious in placing undue weight on outliers. Indeed, it may be desirable in practice to trade off some bias for a smaller variance of the estimators, by trimming extreme weights. We do not, however, have suggestions as to how this might be optimally done. It suggests that in any event the application of the reweighting strategy should be preceded by some consideration of the distribution of the raw values and their weights.

5 South African earnings at the beginning of the post-apartheid era

In order to explore how these techniques work in practice, we applied them to earnings data from the 1994 October Household Survey. This data set ought to be of considerable interest to South African labour economists, since it is the first nationally representative household survey to be conducted in the post-apartheid era. In practice this survey has been difficult to work with, due to data quality issues (Wittenberg 2008b). The income information is particularly problematic, since the national statistical agency converted bracket responses to

[2]These comparisons are not 100% correct since the two sets of estimates were run on different samples. If we reestimate the moments along the lines of Table 3 on the same samples that we ran the EM algorithm on, we get results very similar to those reported in that table.

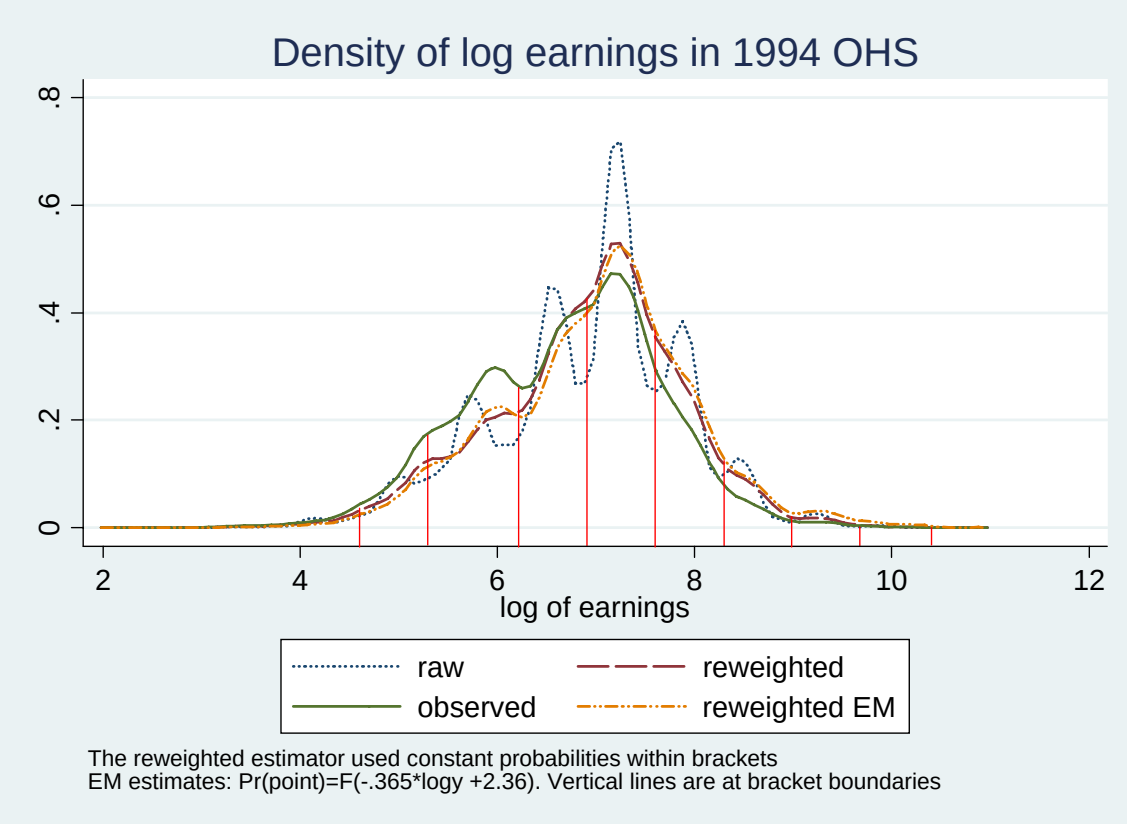


Figure 2: The two "reweighted" density estimates are fairly similar, except that the parametric reweighting by the EM estimates redistributes the density somewhat towards the upper tail.

imputed values without adding an imputation flag to the data. Nevertheless it is possible to separate the point responses from the brackets only responses (Wittenberg 2008a). Implementing this procedure, however, leads to a big loss of observations. Only 44% of responses are valid point values. The remainder are bracket responses. In this case we would expect differences arising from the imputation or reweighting strategies to be magnified, because of the extent of the imputations.

Figure 2 shows standard Epanechnikov kernel density estimates from different approaches. The lines labelled "reweighted" show the results when the two reweighting strategies discussed in this paper are applied to the observed point values. We also show the density $f^{obs}(x)$ as well as a density estimated on the data as supplied by Statistics South Africa, which include the imputed values. The latter shows that the density estimator finds it difficult to smooth over the artificial spikes in the distribution. As far as the observed density is concerned, it is evident that it places too much weight on lower earnings and misses considerable portions of the upper tail. The two "reweighted" estimators provide fairly similar pictures of the density, although the EM estimates result in a noticeable right-ward shift of the distribution.

While the impact of this is not that dramatic in Figure 2 it makes a much bigger difference where the summary statistics are concerned, particularly of the anti-logged figures, i.e. the actual earnings. Table 5 shows that the Statistics South Africa imputations produce a similar mean to the one obtained by imposing constant propensities to report within brackets. These are both significantly higher than the "unweighted" mean[3]. The most startling feature of the table, however, is that imposing a smooth functional form on the propensity to report leads to a 0.084 increase in mean log earnings and a 15.5% increase in mean earnings. This huge increase is driven by two factors. Firstly the "parametric" reweighting procedure changes the relative distribution of weights **within** a bracket. In our case this will lead to a relative downweighting of observations near the lower category boundary and an upweighting towards the right. This leads to a definite increase, although in this case the total impact on mean earnings would only be 1.4%. (Even this might be noteworthy). The main reason for the big increase in average earnings is that the parametric reweighting also

[3]In fact all of the inverse probability weights are multiplied by the Statistics South Africa released person weights, so the "unweighted" figures in Table 5 is in fact a weighted average, using only the Stats SA person weights.

Log Earnings	**unweighted**	**imputed**	**reweighted constant**	**reweighted EM**
Mean	6.699	6.916	6.918	7.002
s.d.	0.964	0.957	0.956	0.982
Median	6.802	7.090	7.026	7.090
Skewness	-0.221	-0.354	-0.342	-0.113
N	12603	28812	12603	12603
Earnings				
Mean	1261.8	1541.3	1550.1	1789.6
s.d.	1562.1	1883.2	1949.3	2603.5
Median	900	1200	1126	1200
Skewness	7.43	8.70	9.10	7.18
N	12603	28812	12603	12603

Table 5: Summary Statistics from OHS94 earnings data

Bracket	Proportion observed	$Pr(g_i = 1\|y_i)$ at midpoint	$Pr(g_i = 1\|y_i)$ cubic model
$1 - 100$	0.58	0.824	0.713
$100 - 200$	0.62	0.702	0.719
$200 - 500$	0.67	0.588	0.617
$500 - 1000$	0.47	0.478	0.480
$1000 - 2000$	0.41	0.379	0.358
$2000 - 4000$	0.35	0.287	0.276
$4000 - 8000$	0.26	0.208	0.249
$8000 - 16000$	0.29	0.143	0.297
$16000 - 33000$	0.41	0.093	0.463
33000+	0.070	0.058	0.764
Notes: $Pr(g_i = 1\|y_i)$ estimated by EM algorithm			

Table 6: Observed and predicted probabilities of giving point answers, OHS94

changes the distribution of weights **across** brackets. This is visible in Figure 2 where the entire distribution in the fourth bracket has shifted to the right. The total density within a particular bracket is not preserved by the parametric procedure. One could force the procedure to maintain the distribution between brackets, by using the parametric procedure only to fix relative weights within a bracket, but this is theoretically unattractive, since it leads to a selection model which is continuous within brackets but discontinuous between them.

The reason why the parametric procedure has such a big impact can be seen in Table 6. It is evident that while there is a clear tendency for the probability of giving point values to decrease with income, the pattern is not monotonic. Modelling the probability of reporting as a simple probit clearly oversimplifies the actual pattern. Two approaches suggest themselves. Firstly, we could estimate the probit model by means of a polynomial in $\log x$. We show in the fourth column of Table 6 what impact this has on the predicted probabilities if we model the relationship with a cubic in $\log x$. In the middle of the distribution the fit is now considerably better, but the estimated probabilities become strange, particularly in the upper tail. If we reestimate the summary statistics with these new weights we get a smaller estimate of average earnings, a figure of $R1588.89$. This is still a 2.5% increase from the simple uniform weighting estimator, although not as large as the 15.5% increase reported earlier. Nevertheless fitting higher and higher polynomials in $\log x$ is essentially *ad hoc*.

A second, more promising, avenue is to add covariates into the selection equation to see whether there are multiple mechanisms at work. Heeringa (1995) suggested for US data that there might be at least two mechanisms accounting for coarsened data: there are recall and competence issues on the one hand; but there are also worries about privacy on the other. The latter mechanism should produce a fairly monotonic pattern in the operation of the coarsening variable. The former, on the other hand, is likely to operate along other dimensions. For instance casual workers and individuals with lower numeracy skills might find it difficult to give precise answers, even though they might be willing to answer the question. In the South African context, Posel and Casale (2005) have suggested that people responding in brackets are more likely to be "proxy

responders", i.e. people providing information about other household members' incomes; people living in large households, where the quality of information about other members may be smaller; White respondents; and high and very low-income earners. The fact that income is relevant even when some of these other variables are controlled for, suggests that the selection mechanism is again non-CAR. Adding covariates into the EM algorithm is theoretically feasible, but practically difficult.

Estimating these type of relationships is beyond the scope of this paper. Nevertheless it is clear that such improved weights would have qualitatively the same impact as the weights derived from the simple probit models estimated here. Compared to the imposition of uniform weights within brackets, this will almost definitely lead to an increase in the estimated average wage, although perhaps not by as much as the 15.5% suggested by our probit with the index linear in $\log x$.

Despite the limitations of imposing uniform weights within brackets, Figure 2 suggests that the simple reweighting estimator performs reasonably well. It also has the advantage that it respects the distribution of observations between brackets. In that sense it is more in keeping with the spirit of nonparametric estimation. Furthermore it is considerably easier to implement. The big difference in Figure 2 is not between the two "reweighting" estimators. It is between these two on the one hand and the observed density and the density including point imputations on the other. Which of the two reweighting strategies one prefers will depend on how one assesses the relative merits of working with a discontinuous selection mechanism versus imposing a continuous one that conflicts with the observed pattern of coarsening.

6 Conclusion

The reweighting strategy works well in estimating density functions and summary statistics in contexts where some of the responses are given only in brackets. It is relatively straightforward to extend these insights to the case where some of the information is completely missing. In that case we will however need a selection mechanism that distinguishes between the different types of coarsening. We leave this to future work. Our Monte Carlo experiments show that that the procedure is successful when the coarsening has happened at random. Indeed the simple uniform reweighting estimator seems to perform adequately even when the underlying process is non-CAR. It is possible to improve on this by estimating the operation of the selection mechanism through our version of the EM algorithm. The Monte Carlo experiment shows that this procedure estimates the selection parameters fairly precisely.

When we apply these procedures to the South African data, we show that the simple reweighting estimator leads to a dramatic increase in mean income (around 23%) when compared to the case where the bracket responses are ignored. If the non-CAR nature of the coarsening is acknowledged there is a further increase in mean income of the order of 2% to as much as 15%. Analysis of this case, however, suggests that a misspecification of the selection mechanism might lead to a reweighting of the data in ways that seem at variance with the empirical distribution across brackets. This is intrinsically little different from the observation that parametrically imputed income values can frequently end up in brackets far removed from where they seem to belong according to the category information (see for instance Posel and Casale 2005). As in the case of our Monte Carlo experiments in section 3.1, such parametric imputations can be extremely useful when the underlying assumptions are close to valid, but they can significantly distort the results if they are not. A similar conclusion seems to apply to the parametrically reweighted density estimates. Use of this procedure would be advised mainly after a thorough analysis of the coarsening process. Even in that case one might want to trim the resulting weights to avoid the excess variability shown in our Monte Carlo simulation.

References

Dempster, A.P., N.M. Laird, and D.B. Rubin, "Maximum Likelihood from Incomplete Data via the EM Algorithm," *Journal of the Royal Statistical Society. Series B (Methodological)*, 1977, *39* (1), 1–38.

Duncan, Greg J. and Graham Kalton, "Issues of Design and Analysis of Surveys across Time," *International Statistical Review*, 1987, *55* (1), 97–117.

Elliot, Dave, *Weighting for non-response: A survey researcher's guide*, London: Office of Population Census and Surveys, Social Survey Division, 1991.

Heeringa, Steven G., "Application of Generalized Iterative Bayesian Simulation Methods to Estimation and Inference for Coarsened Household Income and Asset Data," in "Proceedings of the Survey Research Methods Section, American Statistical Association" 1995, pp. 42–51.

Heitjan, Daniel F. and Donald B. Rubin, "Ignorability and Coarse Data," *The Annals of Statistics*, 1991, *19* (4), 2244–2253.

Holt, D. and D. Elliot, "Methods of Weighting for Unit Nonresponse," *The Statistician*, 1991, *40* (3), 333–342.

Horvitz, D.G. and D.J. Thompson, "A Generalization of Sampling Without Replacement From a Finite Universe," *Journal of the American Statistical Association*, 1952, *47* (260), 663–685.

Kalton, Graham and Ismael Flores-Cervantes, "Weighting Methods," *Journal of Official Statistics*, 2003, *19* (2), 81–97.

Little, Roderick J.A., "Missing Data Adjustments in Large Surveys," *Journal of Business and Economic Statistics*, 1988, *6* (3), 287–296.

Posel, Dorrit and Daniela Casale, "Who replies in brackets and what are the implications for earnings estimates? An analysis of earnings data from South Africa," Paper presented to the Biennial Conference of the Economic Society of South Africa 2005. Available at http://www.essa.org.za/download/2005Conference/Posel.pdf.

Rosenbaum, Paul R. and Donald B. Rubin, "The central role of the propensity score in observational studies for causal effects," *Biometrika*, 1983, *70* (1), 41–55.

Schafer, J.L., *Analysis of Incomplete Multivariate Data*, Boca Raton: Chapman and Hall, 1997.

Wittenberg, Martin, "Income in the October Household Survey 1994," School of Economics and SALDRU, University of Cape Town 2008.

——, "The October Household Survey 1994," School of Economics and SALDRU, University of Cape Town 2008.

Woolard, Ingrid and Chris Woolard, "Earnings inequality in South Africa 1995–2003," Employment, Growth and Development Initiative, Occasional Paper 1, HSRC, Cape Town 2006.

Wooldridge, Jeffrey M., "Inverse probability weighted estimation for general missing data problems," *Journal of Econometrics*, 2007, *141*, 1281–1301.

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