The top tail of South Africa's earnings distribution, 1994-2011

Martin Wittenberg
School of Economics, SALDRU and DataFirst
University of Cape Town
South Africa

Preliminary draft August 2015

1 Introduction

South Africa has long had the reputation of high levels of inequality. Many analysts (Leibbrandt *et al* 2010, van der Berg 2010) concur that inequality in South Africa has not decreased in the post-apartheid era. Most of this literature has focused on the relationship between inequality and poverty. An issue that has received less academic attention is the fate of the relatively affluent. In popular imagery, however, the idea that "the rich get richer" is fuelled both by newspaper reports of conspicuous consumption by the emerging black elite, as well as by the continued privileged position of South Africa's white population. Indeed the idea that South Africa's fat upper tail of the income distribution has fattened is underpinned by reports that incomes in the tail have risen faster than elsewhere (e.g.Blandy 2009).

One of the difficulties in analysing this issue is that South Africa's household surveys are less suited to this purpose than for the analysis of poverty. Firstly, individuals with high incomes are more reticent to divulge their earnings than people with lower incomes. This is reflected in the fact that "bracket responses" are more common at the top end of the income distribution. Frequently bracket responses are given imputed values. Unfortunately this makes the analysis of this part of the income distribution sensitive to the nature of the imputations. This is likely to be particularly problematic for the top income category, where there are no bounds within which to impute. Secondly, information about earnings other than labour income is likely to be poor, so that overall trends in inequality are likely to be understated. Thirdly, refusals to participate in surveys are higher in affluent suburbs than in poor neighbourhoods. The surveys attempt to compensate for this by "weighting" up those respondents that they do find. To the extent to which nonparticipants differ systematically from respondents, the resulting analysis may underestimate inequality.

This paper attempts to analyse the evolution of earnings at the top end of South Africa's income distribution by means of both nonparametric and parametric analyses. One of the key challenges that has to be confronted is in how to combine both the "bracket" and point estimates. Briefly the strategy in the former case is to use assume that the probability of providing a point estimate, conditional on being in a particular bracket, is essentially random. This allows us to reweight the point estimates to account for the individuals replying within brackets. A secondary issue is that we also need to ensure that the weights that we use are consistent between different data sets. To this

end we use Branson's harmonised cross-entropy weights for the South African national data sets. We deflated all datasets post 1995 to October 1995 values, using the CPI, so that all comparisons are to real incomes.

In the parametric analysis we use the Pareto distribution to estimate the Pareto parameter α for all the October Household Surveys and Labour Force Surveys between 1995 and 2007. The Pareto distribution (and various generalisations of it) has been popular in characterising "fat tailed" distributions, i.e. those which can be loosely characterised by a "power law" (cf Mandelbrot). Our justification for using it is based on the initial nonparametric analysis which suggests that tail behaviour can be well approximated by a "power law" as implied by the Pareto distribution.

The Pareto distribution has been used informally in the analysis of South Africa's income distribution. For instance the practice of imputing incomes in South Africa's top income bracket at twice the value of its lower bound is based on a Pareto coefficient estimate of 2 obtained by Charles Simkins (Simkins, personal communication). One of the properties of the Pareto distribution is that $E\left(x|x \geq x_0\right) = \frac{\alpha}{\alpha-1}x_0$, i.e. with $\alpha = 2$ the expected value of incomes in the top bracket would be twice the lower bound x_0.

The only formal use of the Pareto distribution is in a paper by Fedderke *et al* (2004) which attempts to critique estimates of inequality and poverty on the basis of South African household surveys. Unfortunately that analysis is bedevilled by at least two faults. Firstly, it does not seem to deal at all with the issues of incomes reported in brackets. Indeed it seems clear from the authors' discussion of the 1996 October Household Survey (where income was **only** reported in brackets) that the authors merely imputed incomes at the midpoint of each bracket. What they did in the top category is unclear. It is evident that this imputation strategy will heavily affect the parameter estimates. Secondly they used rather generous definitions of "tails", i.e. the threshold above which they estimated the parameter was implausibly low.

Our analysis differs in a further respect from the Fedderke *et al* paper. We analyse the individual earnings distribution, rather than consider the distribution of household (or household *per capita*) income. In particular, we are trying to establish whether there is any evidence that the shape of the tail of the wage distribution has changed.

2 Estimation strategy

2.1 Nonparametric approach

The Pareto distribution is defined by the cumulative distribution function

$$F\left(x\right) = 1 - \left(\frac{x_0}{x}\right)^{\alpha} \tag{1}$$

where x_0 is the cut-off defining the "tail" of the distribution and α is the Pareto parameter. This can be rewritten as the simple power law

$$\log\left(1-p\right) = \alpha \log x_0 - \alpha \log x$$

where p is the cdf evaluated at x. This leads to a simple nonparametric estimation strategy, i.e. graph $\log\left(1-p\right)$ against $\log x$. If the relationship is approximately linear, then a Pareto distribution is a reasonable summary of the shape of the tail distribution. Indeed, one estimation strategy of α is to regress $\log\left(1-p\right)$ against $\log p$. That approach is likely to be less efficient than maximum likelihood, however.

One major obstacle to implementing this nonparametric approach, is that one has to deal with the discretised part of the distribution, i.e. the individuals who report only the bracket into which their income fell. As mentioned above, our approach is to reweight the point estimates within each bracket to deal with the bracket only responses. In essence we estimate

$$P\left(\text{point estimate}|bracket\right)$$

as the (weighted) proportion of individuals within the bracket giving an actual Rand amount. We use the inverse of this probability to reweight the observations.

2.2 Parametric approach

If we only had point estimates then the maximum likelihood estimator is easy to derive from the cdf given in equation 1. The pdf, and hence the contribution to the likelihood is

$$f\left(x\right) = \frac{\alpha x_0^{\alpha}}{x^{\alpha+1}}, \text{ where } x \geq x_0$$

If the probability of giving a bracket response is random within the bracket, then the contribution of these bracket responses to the overall likelihood is simply the probability of falling into the given bracket. This probability for the bracket $[lb_i, ub_i]$ is

$$P\left(x \in [lb_i, ub_i]\right) = \left(\frac{x_0}{lb_i}\right)^{\alpha} - \left(\frac{x_0}{ub_i}\right)^{\alpha}$$

provided that both b_0 and b_1 are above x_0. If we only had observations in our data set guaranteed to be above x_0, this would be easy. Unfortunately we also have to deal with the case where x_0 is in the interior of some bracket $[b_0, b_1]$[1]. In essence we are trying to estimate the parameters of the **conditional** distribution, i.e. $f\left(x|x \geq x_0\right)$ while dealing with observations where it is uncertain whether they belong in this distribution or not. If we knew precisely what the probability was that $P\left(x \in [x_0, b_1] \,|x \in [b_0, b_1]\right)$ we could assume that this fraction of the occurrence of the sample observations of the interval $[b_0, b_1]$ belonged in our estimation sample while the rest did not. We can achieve the same purpose by "downweighting" observations corresponding to such brackets by the proportion $P\left(x \in [x_0, b_1] \,|x \in [b_0, b_1]\right)$. We estimate this proportion on observations for which we have point information, i.e. we calculate the (weighted) proportion of individuals within the bracket $[b_0, b_1]$ which fall above the cut-off x_0. Note that this strategy is available only in cases where we have some point information on the bracket which straddles the cut-off. In particular it cannot be used where we **only** have bracket information. In this case we fall back on the strategy of estimating the parameters only for tail distributions where the cut-off coincides with a bracket boundary.

Numerically the downweighting procedure is equivalent to multiplying the individual contribution to the log-likelihood by $P\left(x \in [x_0, b_1] \,|x \in [b_0, b_1]\right)$. Putting this all together, the log likelihood

[1] This raises the obvious question as to why we don't insist on estimating the parameters only at points which fall on bracket boundaries. There are two responses to this. Firstly, in one of the datasets (1995) the brackets are different for different categories of responses, so that there is no single cut-off which would not fall into the middle of some or other category. Secondly we want to keep the analyses comparable across years, so we need to fix the cutoff in real terms. With fixed categories this implies that even if at one point in time the cut-off coincides with a category boundary, it will not do so in subsequent years.

for the sample will be given by

$$\ln L\left(\mathbf{x}|\alpha\right)=\sum_{d=1}\ln\left\{\frac{\alpha x_0^{\alpha}}{x_i^{\alpha+1}}\right\}+\sum_{\substack{d-0\\ x\in[lb_i,ub_i]\\ lb_i\geq x_0}}\ln\left\{\left(\frac{x_0}{lb_i}\right)^{\alpha}-\left(\frac{x_0}{ub_i}\right)^{\alpha}\right\}+\sum_{\substack{d-0\\ x\in[lb_i,ub_i]\\ lb_i<x_0<ub_i}}P\left(x\geq x_0|lb_i<x<ub_i\right)\ln\left\{1-\left(\frac{x_0}{ub_i}\right)^{\alpha}\right\} \tag{2}$$

where d is an indicator variable indicating whether the individual provided an actual amount.

An additional complication arises in the estimation of this model, given that there was differential response. The underrepresentation of white South Africans in the national surveys is likely to be particularly problematic when dealing with the top incomes. There is little option but to use the sample design weights adjusted for nonresponse. In effect this means that the actual estimation procedure is a pseudo-maximum likelihood one, i.e. we assume that the population moment condition

$$E\left[\frac{\partial \ln L}{\partial \alpha}\right]=0$$

can be consistently estimated by the weighted sample moment condition

$$\sum_i w_i\frac{\partial \ln L_i}{\partial \alpha}=0$$

3 The data

We use all the October Household Surveys from 1995 onwards and all waves of the Labour Force Surveys until September 2007. There are a number of complications in putting these datasets next to each other. The first of these is that there was a change in measurement between the October Household Surveys and the Labour Force Surveys. In the former, there were two questions asked in relation to earnings: one for wage earners and one for earnings from self-employment. In the Labour Force Surveys there is only one earnings question. Combining information from the two sources is, in principle, not difficult. It does rather complicate the bracket part of the analysis, because of the permutations involved. This becomes crippling for the nonparametric part of the analysis. Consequently we are presenting information only for wage information for the October Household Surveys, while all earnings are considered for the LFSs.

There are also difficulties in particular surveys. The questionnaire in the 1995 OHS is particularly unfortunate, because the respondent was first asked to select one of a range of brackets and then to specify whether the bracket information pertained to daily income, weekly, monthly or annual. So when all incomes are converted to a monthly reporting period, one has different brackets for daily, weekly, monthly and annually paid workers. This problem is exacerbated by the fact that the first bracket is overly big, i.e. R1 to R999. A daily paid worker indicating this bracket would get an equivalent monthly bracket ranging from R22 to R21978. This will definitely reach into the top tail of the income distribution, although it is quite unlikely that most daily paid workers would actually sit near the top end of that bracket.

4 Results

The results from the nonparametric analysis are shown in Figures 1 and 2. Figure 1 shows the entire top tail. The lower part of that figure is very noisy due to the fact that it is heavily influenced

by a few outliers. The September 2005 and September 2000 are particularly anomalous, because the maximum in those years was particularly high. In Figure 2 we concentrate on the distribution between R4000 and around R22000 per month (in 1995 Rands). This range encompasses 98% of all observations in the tail.This diagram is quite striking in that most of the tail distributions lie almost on top of each other. There are two clearly anomalous distributions: that for October 1995, which was most egalitarian (within the tail) and that for October 1999 (most inegalitarian). There is no clearly discernible pattern for the rest. It is not the case that the tail has shifted out monotonically.

In both Figures we have added in the regression line which was estimated from a pooled regression (across all surveys) of $\log(1-p)$ on $\log(w)$. The estimate of α derived in this way is 1.77.

Figure 2, in particular, suggests that a Pareto distribution is likely to fit the data fairly well. The detailed estimation results are given in Table 1. In these tables we have ensured that the cut-offs correspond to a category boundary in the 1996 OHS, since that dataset only has categorical information. Because the category boundaries remained fixed in nominal terms for all subsequent years, we inflated the cut-off in line with changes in the CPI. This is equivalent to maintaining a fixed real cut-off and deflating the income values. In order to test the sensitivity of our results, we tried three different cut-offs: R4501, R6001 and R8001. To illustrate what happens when an inappropriately small cut-off is selected, we also report the results for a cut-off of R2501.

Looking across the columns of the table, there is little evidence that the estimates are sensitive to the choice of the cut-off, except if the low value of R2501 is selected. Furthermore there is little evidence, going down the columns, that there has been any clear evidence for systematic changes over time. Given the nonparametric evidence shown earlier, this is hardly surprising.

The "pooled" estimates of α are around 1.86, which agrees reasonably with the slope of -1.77 that was estimated for the results in Figure 1.

5 Conclusion

There is no evidence in this analysis which suggests any tendency for the tail of the income distribution to become fatter. If anything the results seem remarkably stable. It is, of course, possible that a shifting out of the distribution has been exactly offset by increasingly systematic measurement bias (e.g. through nonresponse). However that is not a particularly parsimonious explanation of the lack of a trend.

There is also no evidence to back up the assertion of the Fedderke *et al* paper that one can draw vastly different conclusions from different datasets. There are certainly a number of anomalies. We have pointed out the peculiar nature of the October 1995 and October 1999 distributions. Similarly it is clear that there are many practical difficulties in the way of doing decent analyses of even the simple type attempted in this paper. The fact that many individuals report incomes in brackets is a nontrivial issue that needs to be confronted. Nevertheless it is possible to get meaningful information with some additional assumptions.

More substantively, a Pareto coefficient of around 1.8 suggests that the distribution is definitely fat-tailed. Interpreted literally, it would suggest that the distribution has a mean, but no variance. In essence the probability of observing extreme values does not die out sufficiently rapidly for the variance to remain bounded. It is statistical reflection of the casual observation that there are quite a lot of filthy rich South Africans. The existence of this tail may very well give rise to the perception of the "rich getting richer" which is the subject of South African dinner tales.

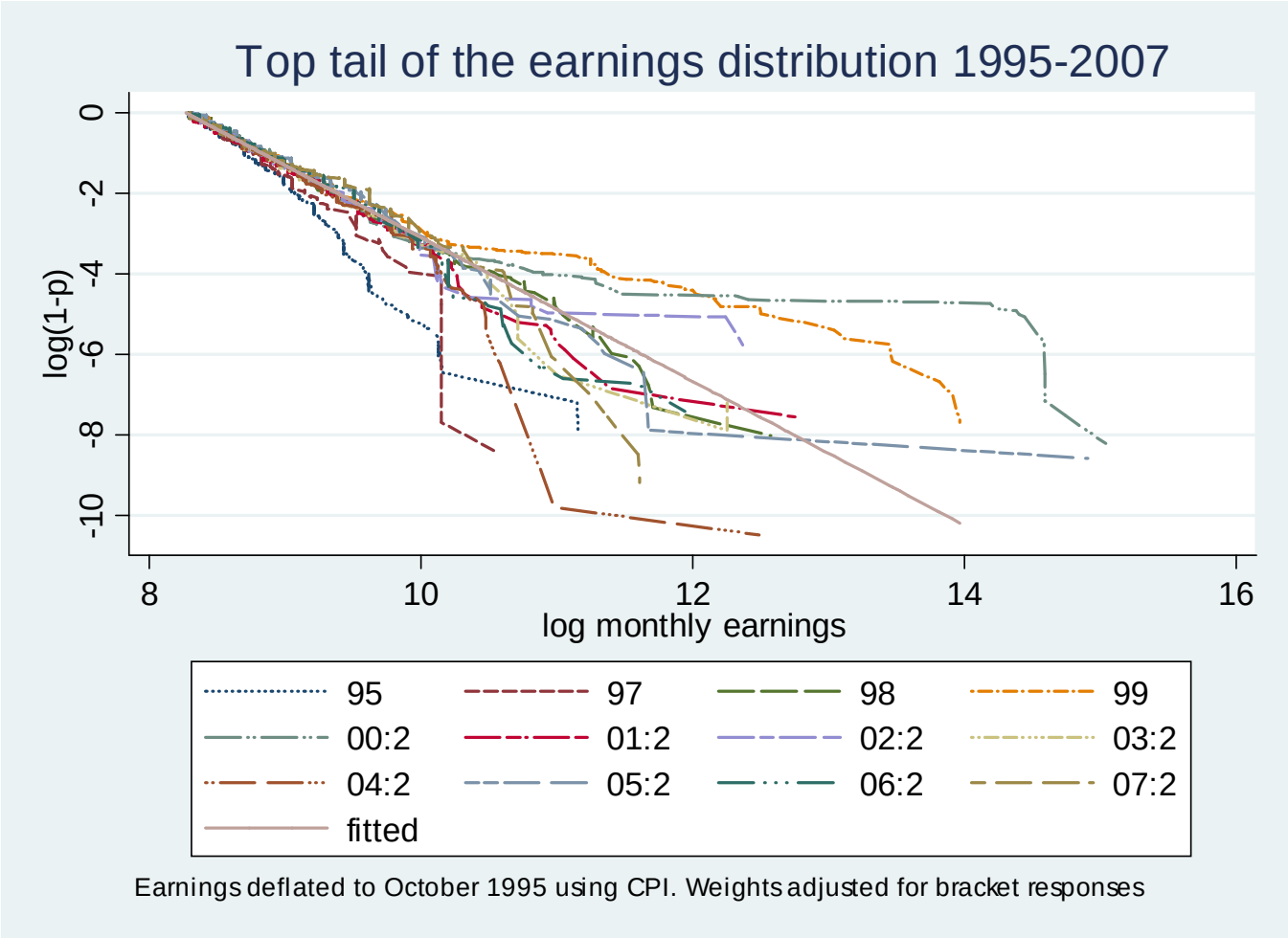

Figure 1: Incomes above R4000 (in 1995 value). Graph of $\log(1-p)$ against $\log w$.

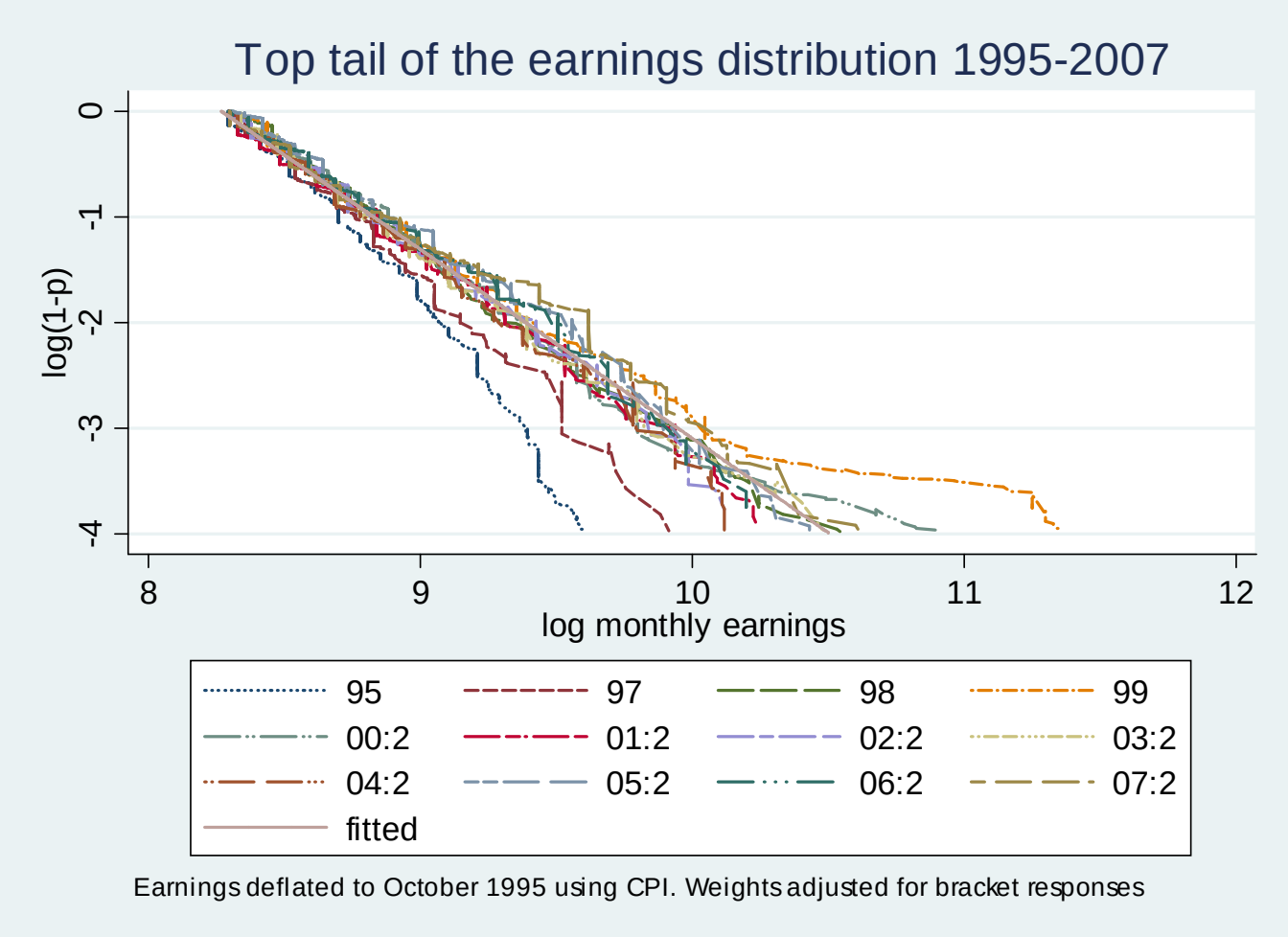

Figure 2: Incomes from R4000 to R22000 per month (1995 values). $1 - e^{-4} = 0.982$ of all incomes above the cutoff are shown