The top tail of South Africa's earnings distribution, 1993-2014: Evidence from the Pareto distribution*

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

Note: A revised form of this paper will be released as a REDI3x3 working paper. This version should not be cited.

1 Introduction

South Africa has long had the reputation of high levels of inequality. Many analysts (Bhorat, Van Der Westhuizen and Jacobs 2009, Leibbrandt, Woolard, Finn and Argent 2010, Leibbrandt, Finn and Woolard 2012, van der Berg 2011) concur that inequality in South Africa has not decreased in the post-apartheid era. Much 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" (Blandy 2009) 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.

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.

*Funding from REDI3x3 for this project is gratefully acknowledged. Much earlier versions of this paper were presented at seminars at the University of Michigan and at a Data Quality workshop hosted by DataFirst in Cape Town.

An additional issue that was flagged by Burger and Yu (2007) and Wittenberg (2014b, 2014a) is the problem of extreme values, many of which seem outright data errors. Cowell and Flachaire (2007) have noted that the reliable estimation of inequality measures needs to confront the dual issue of data contamination in the top tail as well as the sparseness of data points. They note that the top tails of income distributions tend to be "heavy-tailed" and random samples tend to underestimate the true weight in the upper tail. Consequently they suggest that the estimation of inequality should combine the typical nonparametric estimates (i.e. measuring inequality purely on the empirical distribution function) with a parametric estimation of the contribution of the top tail. This they do by using the Pareto distribution.

Our contribution in this paper is to estimate the parameters of a Pareto distribution of the distribution of earnings as measured in the Post-Apartheid Labour Market Series (PALMS) dataset (Kerr, Lam and Wittenberg 2016) and to assess how much the robust estimation technique of Cowell and Flachaire alters our understanding of earnings inequality. Neither of these have been done properly on South African survey data. In the process we will also highlight, yet again, the importance of measurement issues for the understanding of the trends. The paper also makes a number of methodological contributions. For instance we show how the Pareto distribution can be used to flag outliers.

The estimated parameters of the Pareto distribution are interesting in their own right. Mandelbrot (1960) has suggested that heavy-tailed distributions of the Pareto type are "stable" in the same sense that Gaussian distributions are, i.e. they are the sum of shocks each distributed also as Pareto-type. These distributions are likely to describe the data well in contexts where the outcome is the result of a small number of "big" shocks. Arguably earnings distributions fit this description, since increases due to promotions and changes in jobs tend to have a larger impact than incremental wage adjustments within a job category. In the case of self-employment income, where windfall gains are possible, this is even more likely to be the case. The size of the Pareto parameter is therefore useful as a measure of how large the upper tail is.

We begin our discussion with a brief review of the literature and of the data that we will be using. We then present our estimation strategy. This starts with a non-parametric view of the top tail of the distribution and continues to discuss three methods of estimating the Pareto parameter. We use a pseudo-maximum likelihood technique in the rest of the paper. We turn to discuss the problem of outlier detection in the context of estimating Pareto parameters and then present the Cowell-Flachaire (2007) procedure. The results and discussion round off the paper.

2 Literature Review

Earnings inequality in South Africa has been discussed in a number of papers (Leite, McKinley and Osorio 2006, Heap 2009, Tregenna 2011, Tregenna and Tsela 2012). Wittenberg (2014b, 2014a) suggests that earnings inequality has increased over the post-apartheid period, but notes that measurement issues affect the reliability of these estimates. Some of the key points are:

- Changes in the instrument and sampling procedures:

  The October Household Surveys undersampled small households (Kerr and Wittenberg 2015) and as a result probably underenumerated certain types of workers (e.g. domestic workers living in the backrooms of their employers' homes). This problem is compounded by the fact that the LFSs found many more informal sector workers, particularly in agriculture (Neyens

and Wittenberg 2016). This means a sharp disjuncture between the OHS and LFS wage and employment series, particularly in relation to self-employed workers.

- Extreme values

  Burger and Yu (2007) commented on the fact that some datasets seemed to contain many more "millionaires" than others. Wittenberg (2014b, 2014a) discusses several possible "outlier detection" routines and shows that removing the extreme values has an appreciable effect on the mean of realearnings. His preferred method is based on a Mincerian regression. Observations with extreme standardised residuals (more than an absolute value of 5) are marked as outliers.

- Bracket responses

  Respondents that were unwilling to disclose a Rand earnings amount were given the option of specifying a range instead. Wittenberg (2014b) provides evidence that individuals who responded in brackets were more likely to be high earners. He also shows that imputing midpoints or means of the ranges (as much of the other literature does) is likely to distort both the estimate of the mean and of inequality measures. Instead he suggests that the bracket information can be used to reweight the point responses. Alternatively he suggests a multiple stochastic imputation routine.

- Missing values

  There are a number of respondents in the pre-QLFS surveys who supply neither Rand nor bracket information. Again these seem to be predominantly high income individuals. Wittenberg (2014b, 2014a) argues that a multiple imputation routine provides the best way of dealing with these cases.

None of these approaches deals well with the situation of data contamination of a "heavy-tailed" distribution, such as the Pareto. The criterion for judging extreme values is based on what looks "extreme" in the context of a normal distribution. However this will lead to an over-rejection of observations that might well look less extreme if the true distribution is Pareto. Secondly the multiple imputation routine is a version of a "hot deck", i.e. it is a draw from the empirical distribution function. This means that if there are too few high values to begin with (or if these have been over-zealously removed in the outlier detection routine) then they will not be created in the imputation. In this context the Cowell and Flachaire (2007) procedure looks more promising, since it takes the possibility of a heavy tail seriously.

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). The Pareto distribution has been used formally in a paper by Fedderke, Manga and Pirouz (2004) which attempts to critique estimates of inequality and poverty on the basis of South African household surveys. Unfortunately that analysis is bedevilled by several faults. Firstly, the authors attempt to calculate per capita incomes on data which are not really suited to that analysis. In particular the OHSs that are not linked to income and expenditure surveys provide information on labour market earnings, but not on other types of income. Secondly, in so far as the labour earnings are utilised, the analysis 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. Thirdly they used rather generous definitions of "tails", i.e. the threshold above which they estimated the parameter was probably too low, as our analysis below will suggest.

More recently Alvaredo and Atkinson (2010) have investigated the top tail of South Africa's wealth distribution using tax data and estimated various Pareto coefficients in the process. These coefficients were estimated from the income shares of groups (the top 1% and top 0.1%) and not off microdata.

3 The Data

We make use of the Post-Apartheid Labour Market Series (PALMS), version 3.1 (Kerr et al. 2016). This dataset combines the labour market information from the Project for Statistics on Living Standards and Development (1993), the October Household Surveys from 1994 to 1999, the biannual Labour Force Surveys from February 2000 to September 2007, and the Quarterly Labour Force Surveys from 2008 through to the fourth quarter of 2015. Earnings figures were not released with the QLFSs in 2008 and 2009, nor were the 2015 ones available at the time of doing this research. The earnings information in 1996 was collected exclusively in brackets and consequently we excluded that survey. We therefore have usable information from 42 separate surveys. Given the fact that we will restrict our analyses to individuals earning more than R6000 per month (in real June 2000 values) we end up with around 75 000 individually reported incomes.

The PALMS dataset provides several useful tools for analysing earnings across time. Firstly it has attempted to harmonise definitions. Secondly, it provides a set of harmonised sampling weights to ensure that shifts in the dataset are not due to simple shifts in the demographic models that underpin the weights (Branson and Wittenberg 2014). Thirdly it calculates a set of "bracketweights" which can be used to reweight the reported Rand incomes to account for individuals who responded in brackets (Wittenberg 2014b). Fourthly it marks extreme values using the standardised residuals from a Mincerian regression as diagnostics (Wittenberg 2014b). In this paper we do not use the multiple imputations also released with PALMS, since it isn't clear that the methodology is consistent with the attempt to measure the Pareto coefficient.

For the purposes of this paper we excluded all self-employed agricultural workers, since they are measured inconsistently across time. The number of this type of worker increases by over a million between Octtober 1999 and February 2000. Furthermore at the end of the LFS period, this type of employment almost vanishes again in the survey data (Neyens and Wittenberg 2016).

4 Methods

4.1 Properties of the Pareto distribution

The Pareto distribution is defined by the cumulative distribution function

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

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

$$\log(1-p) = \alpha \log x_0 - \alpha \log x \tag{2}$$

where p is the cdf evaluated at x.

The pdf of the Pareto is

$$f(x) = \frac{\alpha x_0^\alpha}{x^{\alpha+1}}, \text{ where } x \geq x_0 \tag{3}$$

and it follows that

$$E(x|x \geq x_0) = \frac{\alpha}{\alpha - 1} x_0 \tag{4}$$

Note that this is defined only if $\alpha > 1$. It is easy to verify that this relationship holds for any cut-points higher than x_0, i.e. $E(x|x \geq x') = \frac{\alpha}{\alpha-1}x'$ for any $x' \geq x_0$. It is worth noting that the variance of the Pareto distribution is defined only if $\alpha > 2$. It is in this sense that it is potentially "heavy-tailed" - extreme outcomes happen with a sufficiently high probability that the variance can be undefined.

The quantile function is given by

$$Q(F; q) = \frac{x_0}{(1-q)^{1/\alpha}}$$

Combining the previous two results – noting that the total income of the top 1% is the conditional mean above $Q(F; 0.99)$ multiplied by the population – it is straightforward to show that the ratio S_1 of the share of the top 1% to the ratio $S_{0.1}$ of the top 0.1% will be given by

$$\frac{S_1}{S_{0.1}} = 10^{1-\frac{1}{\alpha}} \tag{5}$$

4.2 Estimation strategies

The equations provided above provide at least four different *parametric* approaches to estimating α, but they also suggest a simple nonparametric sense-check of the data.

4.2.1 Nonparametric approach

The power law formulation in equation 2 is useful as a starting point, since it provides a simple visual check on whether the parametric approach is sensible or not. We graph $\log(1-p)$ against $\log x$. If the relationship is approximately linear, then a Pareto distribution is a reasonable summary of the shape of the tail distribution.

4.2.2 Regression

The power law equation can also be used to estimate α, i.e. we regress $\log(1-p)$ on $\log(x)$. We suspect that this is the approach taken by Fedderke et al. (2004). This approach is likely to be less efficient than the pseudo-maximum likelihood version that we will adopt.

4.2.3 Method of moments

The conditional moment equation (4) can be used to define a method of moments estimator, i.e.

$$\widehat{\alpha}_{MoM} = \frac{\overline{x}}{\overline{x} - x_0}$$

where $\overline{x}$ is the sample mean in the top tail. Monte Carlo simulation studies (available from the author) suggest that this procedure is likely to be considerably less efficient than maximum likelihood.

4.2.4 Share of the top 0.1% within the top 1%

The ratio of the shares given by equation 5 is used by Alvaredo and Atkinson (2010) to estimate the Pareto coefficient on tax data, i.e.

$$\widehat{\alpha}_{share} = \frac{1}{1 - \log_{10}(S_1/S_{0.1})}$$

This requires knowing how much of the total income goes to the top 1% and the top 0.1% respectively. Since S_1 and $S_{0.1}$ are derived from the conditional means above $Q(F; 0.99)$ and $Q(F; 0.999)$ respectively this is an estimator that depends on the ratio of two sample moments far up the distribution and is likely to be noisy on survey data.

4.2.5 Maximum likelihood

We can use the pdf given in equation 3 to derive the maximum likelihood estimator

$$\widehat{\alpha}_{ML} = \frac{1}{\frac{1}{n}\sum \ln x_i - \ln x_0}$$

This is related to Hill's estimator of the rate of decrease of the distribution function in the tail (Hill 1975). His estimator is

$$\widehat{\alpha} = \frac{1}{\frac{1}{r}\sum_{i=1}^{r} \ln y^{(i)} - \ln y^{(r)}}$$

where $y^{(1)}, y^{(2)}, ..., y^{(r)}$ are the r largest values ranked from largest downwards. The main difference between our approaches is that we effectively take a fixed cut-off x_0, whereas the Hill estimator allows it to vary with the data. We show below some sensitivity analyses in which we vary x_0, which is akin to varying r, which is normally taken to be some fixed proportion of the sample (Cowell and Flachaire, for example, use $n/10$).

The reason why we prefer a fixed cut-point (corresponding to a fixed level of realearnings) is that it allows us to more effectively compare what happens to the tail distribution over time. Furthermore, as we show below, this allows us to check for the presence of outliers among the largest values. The Hill procedure, by contrast, needs to assume that everyone of the r largest values is measured correctly.

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. This problem is exacerbated by the fact that we do a second level reweighting to account for individuals who gave bracket responses. 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

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

where w_i are the sample weights adjusted for bracket response.

We implement the estimation procedure using Stata's maximum likelihood routine, which allows us not only to weight the data but to calculate standard errors robust to clustering. These standard errors are markedly bigger than they would have been under the assumption of independent sampling.

4.3 Dealing with Outliers

A key question for the empirical analysis is how to flag outliers without relying on criteria derived from the normal distribution for judging which observations are extreme. It is useful to rehearse some of the standard approaches used in the outlier detection literature (see (Billor, Hadi and Velleman 2000) for a review). One simple approach, adopted, for instance by Cowell and Flachaire (2007), is to successively delete each observation and see the impact this has on the parameter estimates. Observations that have a disproportionate impact can be flagged as problematic. The problem is that this doesn't deal with the potential that data contamination may involve a *cluster* of problematic observations. Indeed the empirical work in Wittenberg (2014b) suggests that this is often the case. More generally the problem is that the presence of outliers will contaminate any statistics calculated to detect those outliers. Consequently a standard approach is to begin with a small subset of observations assumed safe from contamination and then to add in observations that are deemed to also be "safe" given the empirical information in the safe set. The BACON algorithm (Billor et al. 2000), for instance, judges observations to be safe based on their Mahalanobis distance, i.e. $\sqrt{(\mathbf{x}_i - \overline{\mathbf{x}})' \mathbf{S}^{-1} (\mathbf{x}_i - \overline{\mathbf{x}})}$, from the safe set, where $\overline{\mathbf{x}}$ and $\mathbf{S}$ are the mean and covariance matrix calculated on the safe set and $\mathbf{x}_i$ is the vector under consideration. In the case of a univariate distribution this measure is just $\frac{|x_i - \overline{x}|}{s}$, i.e. it is akin to a t-statistic evaluating the probability of observing an observation of size x_i (or more extreme) given that the true mean is $\overline{x}$. Indeed these statistics are compared to a t or normal distribution to assess the probability of the observation coming from the same distribution.

This test statistic will not work in the case of a Pareto distribution for two reasons. Firstly in many cases (including the South African one as we will show) the Pareto parameter is in the range where the variance is not defined, so that asymptotically the t-statistic does not exist. Secondly given the focus of the Pareto distribution on the upper tail, the "safe sample" will be asymmetrically defined and not picked around the median of the distribution. This means that $\overline{x}$ from the safe sample will not be a reasonable or consistent estimate of the population mean.

Our procedure starts from the assumption that the k smallest observations $x_{(1)}, x_{(2)}, \ldots, x_{(k)}$ just above the cut-off x_0 constitute the safe sample. We pick $k = 100$, which is small in the context of the typical sample size in the top tail. We then obtain an initial estimate of the Pareto parameter $\widehat{\alpha}_1$ on the safe sample. We then calculate the probability of observing the observations $x_{(k+1)}, x_{(k+2)}, \ldots, x_{(n)}$ (or ones more extreme) on the assumption that the distribution is truly Pareto with parameter $\widehat{\alpha}_1$. The probability of observing $x_{(j)}$ or values higher than it will be given by

$$P\left(X \geq x_{(j)}\right) = \left(\frac{x_0}{x_{(j)}}\right)^{\widehat{\alpha}_1} \tag{6}$$

This probability can be compared to the empirical distribution function. Assume that w_j is the weight of observation $x_{(j)}$ and define the empirical cumulative weight $W_j = \sum_{i=1}^{j} w_{(i)}$. Then the

empirical estimate of $P\left(X \geq x_{(j)}\right)$ assuming that $x_{(j)}$ is the *last* properly measured observation is

$$\widehat{p}\left(X \geq x_{(j)}\right) = \frac{w_j}{W_j} \tag{7}$$

We can compare the theoretical probability $P\left(X \geq x_{(j)}\right)$ to the empirical one $\widehat{p}\left(X \geq x_{(j)}\right)$. If the ratio is too small we would reject the idea that $x_{(j)}$ forms part of the same distribution as the safe sample. Our criterion is

$$\text{accept } x_{(j)} \text{ into the safe sample if, and only if, } \frac{P\left(X \geq x_{(j)}\right)}{\widehat{p}\left(X \geq x_{(j)}\right)} \geq \tau \tag{8}$$

The constant τ determines how easily the procedure accepts extreme values. Values of τ close to one will be less forgiving than values closer to zero. For our empirical analysis we used $\tau = \frac{1}{2}$. After the first iteration of the procedure a number of additional points will be flagged as part of the "safe sample". This new safe sample is then used to re-estimate the Pareto parameter to yield $\widehat{\alpha}_2$. The probability of observing points $x_{(j)}$ outside the safe sample are then recalculated according to formula 6 (with $\widehat{\alpha}_2$ rather than $\widehat{\alpha}_1$) and again compared to the empirical probabilities (equation 7) which may lead to yet further additions to the safe sample. The procedure terminates if either the entire sample is marked safe or the observations outside the safe sample have a much lower probability of occurring than their weight in the sample suggests. The final split into safe sample and outliers is internally consistent, in the sense that the Pareto parameter estimation is not contaminated by the outliers and the outliers look anomalous in light of the Pareto coefficient.

It should be noted that this outlier detection procedure will only capture anomalous observations outside the range of the "safe" values, i.e. if there is a data capture error (e.g. shifting the decimal point two places) which does not, however, move the observation far out into the top tail, it will not be caught by this procedure. This, of course, is equally true of most other univariate outlier detection algorithms. Furthermore unlike the regression procedure it does not take into account the values of any covariates. Lastly this procedure treats errors in the earnings distribution asymmetrically: implausibly large values will be marked as dubious and excluded from the analysis, while implausibly small ones will escape such scrutiny. Since we are less concerned about the bottom of the distribution this does not concern us here, but it would raise questions in a context where we want to investigate characteristics of the distribution as a whole.

4.4 Smoothing the estimation of the Pareto parameter

The estimation procedure outlined in the previous section looks at extreme values only in the context of one particular survey. Nevertheless in PALMS we have over fifty surveys, with earnings information in 42 of them. What may look anomalous in one survey may look less so when compared to adjoining periods. For this reason we also pool surveys within 8 quarters of a particular period and run the Pareto estimation/outlier detection algorithm outlined in the previous section on that pooled sample.

4.5 The Cowell-Flachaire procedure for robust estimation of means and inequality

Cowell and Flachaire (2007) argue that the potential of data contamination together with the fact that surveys are likely to underestimate the true importance of the top tail necessitate the use of

hybrid estimation techniques. In particular they suggest that the distribution should be split into two: the top $(100p_{tail})\%$ and the bottom. Within the bottom part of the income distribution one would use the standard nonparametric estimation techniques, i.e. calculate the mean and inequality measures using the empirical distribution function. In the top part, however, one uses parametric estimates. More concretely, the population mean would be estimated as

$$\widehat{\mu} = (1 - p_{tail})\,\widehat{\mu}_0 + p_{tail}\widehat{\mu}_{tail}$$

The mean in the bulk of the distribution $\widehat{\mu}_0$ would be calculated in the usual way, but the mean of the upper tail $\widehat{\mu}_{tail}$ would be estimated as$\frac{\widehat{\alpha}}{\widehat{\alpha}-1}x^*$ (see equation 4) where x^* is the lower bound of the upper tail as defined by the fraction p_{tail}. Effectively this discards the top $p_{tail}n$ observations and replaces them with the parametric estimate.

Cowell and Flachaire make the point that p_{tail} should be selected much smaller than the number of observations on which the Pareto parameter is estimated. Furthermore it should be selected so that $p \to 0$ as $n \to \infty$ to ensure consistency. In their analysis they pick $p_{tail} = 0.04 * n^{-\frac{1}{2}}$ which means that effectively $0.04 * n^{\frac{1}{2}}$ observations are not used in the "nonparametric" part of the estimation. In the case of PALMS this is a handful of observations per survey.

Cowell and Flachaire provide formula for the Generalised Entropy or Atkinson inequality measures using the same general approach.

It can be shown that the Gini coefficient will be given by

$$Gini = (1 - p_{tail})\,(1 - s_{tail})\,G_0 + s_{tail} - p_{tail} + p_{tail}s_{tail}G_{tail}$$

where s_{tail} is the share of total income accruing to the top $(100p)\%$, G_0 is the Gini coefficient estimated nonparametrically on the bottom part of the distribution and $G_{tail} = \frac{1}{2\alpha-1}$ which would again be estimated using the Pareto coefficient. The total income accruing to the top tail is $\frac{\alpha}{\alpha-1}x^*p_{tail}N$ where N is the total population size. The total accruing to the bottom would be estimated in the standard way as $(1 - p_{tail})\,\widehat{\mu}_0 N$. The share s_{tail} can therefore be estimated as the ratio of $\frac{\alpha}{\alpha-1}x^*p_{tail}$ to $\widehat{\mu}$.

5 Results

5.1 Nonparametric Analysis

Our first look at the data is provided by 1. The graph represents information from surveys two years apart. Several features are apparent in this diagram. Firstly many of these trajectories resemble straight lines for the bulk of the distribution, but change tack in the last few observations. Secondly we see that some surveys have a markedly different trajectory from the others. The October 1999 Household Survey is particularly noteworthy in this regard, but the third quarter of QLFS 2012 and the the third quarter of 2014 also look anomalous. Despite these problems linearity does not seem far-fetched and so we turn to parametric estimates.

5.2 Parametric estimates: regression, method of moments and maximum likelihood

Figure 2 presents three different approaches to the estimation of the Pareto coefficient discussed above. The left-hand side panel presents the regression and method of moments estimates, while

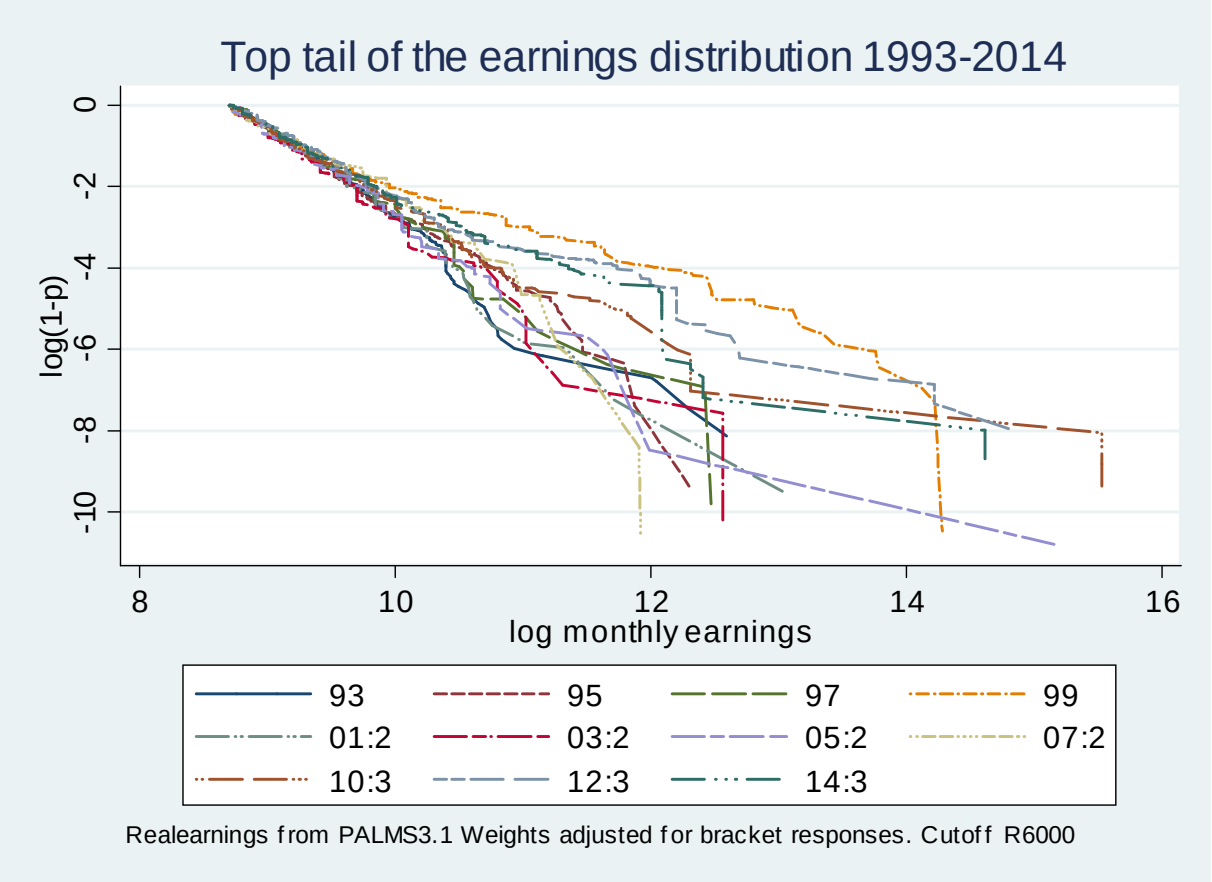


Figure 1: Graph of $\log(1-p)$ against $\log(x)$ where $p = P(X \geq x)$ and x is real earnings.

the right hand panel shows the pseudo-maximum likelihood results together with a robust 95% confidence interval. It is evident that the regression and method of moment estimates are more volatile, particularly when compared to the range of the confidence interval. To interpret these results it should be remembered that lower values correspond to much higher levels of inequality. Distributions with Pareto coefficients below 2 are so thick-tailed that they do not have a variance, and most surveys end up giving point estimates in this range. It is also noticeable that there seem to be major changes in the size of the coefficient over implausible short time horizons. In particular the big reversal between October 1999 and February 2000 is astonishing.

The results of these preliminary analyses confirm our *a priori* assumptions that the pseudo maximum-likelihood approach will be the most reliable available, particularly when combined with the robust estimation of confidence intervals. There are also ample indications that outliers and measurement issues will be important.

5.3 The importance of the cut–offs

It is important to pick a value of the cut-off above which the Pareto coefficient remains relatively stable. In Figure 3 we present some evidence on how sensitive the results are to different choices of the boundary between the "top tail" and the rest. It appears that with the lower cut-off of R4500 per month the estimated Pareto coefficient is noticeably lower in virtually all time periods. The cut-off of R6000 provides lower Pareto estimates in the earl 2000s when compared to higher cut-offs, but this is not a consistent pattern. Given that there is a trade-off between the size of the sample on which the coefficient is estimated and the stability of the parameter, we thought that there was little stronger evidence in favour of going to yet a higher value.

5.4 Outlier detection

Given the results from the literature (Burger and Yu 2007, Wittenberg 2014b) and the noise evidenced in Figure 1 it is not surprising that outliers mattter for the results. We adopted three approaches to outlier detection. Firstly we used the regression approach implemented in PALMS and discussed in Wittenberg (2014b) This approach resulted in the removal of 345 observations (across all waves) from the analysis. Secondly we flagged observations as outliers based on the iterative procedure described in Section 4.3. This procedure led to the removal of only 61 observations. Interestingly enough four of these were not flagged by the regression routine, because key control variables were missing. The third approach was the smoothed approach utilising adjoining datasets, as described in Section 4.4. That approach was even more conservative, flagging only twenty-six observations as outliers.

The impact of the difference between the regression and the Pareto approach can be seen in Figure 4. Both panels should be compared to the "raw" distribution shown in Figure 1. It is evident that the most egregious "zig-zags" in the distribution have been eliminated. Nevertheless it is also clear that the regression approach has cleaned out high values more aggressively than the Pareto approach developed in this paper. It is worth noting, since this will be of some importance in the discussion later, that the October 1999 trajectory on the extreme right of both diagrams has been aggressively pruned by the regression approach, whereas it is largely intact on the Pareto approach. Indeed that is hardly surprising given that it approximates a straight line in this log-log space reasonably well.

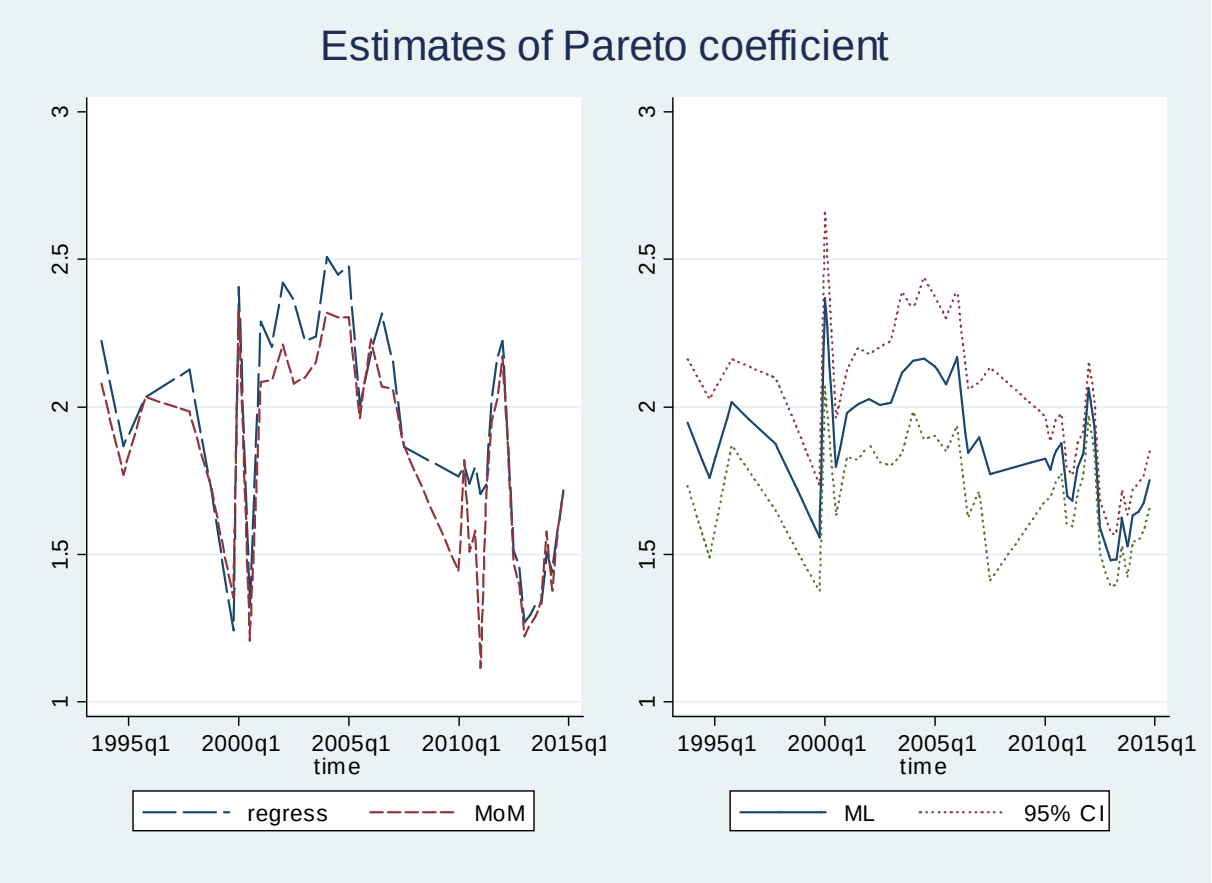


Figure 2: Estimates of the Pareto coefficient according to three different estimation techniques. Real earnings above R6000 per month (June 2000). 95% confidence interval for ML estimates computed with standard errors robust to clustering on primary sampling unit.

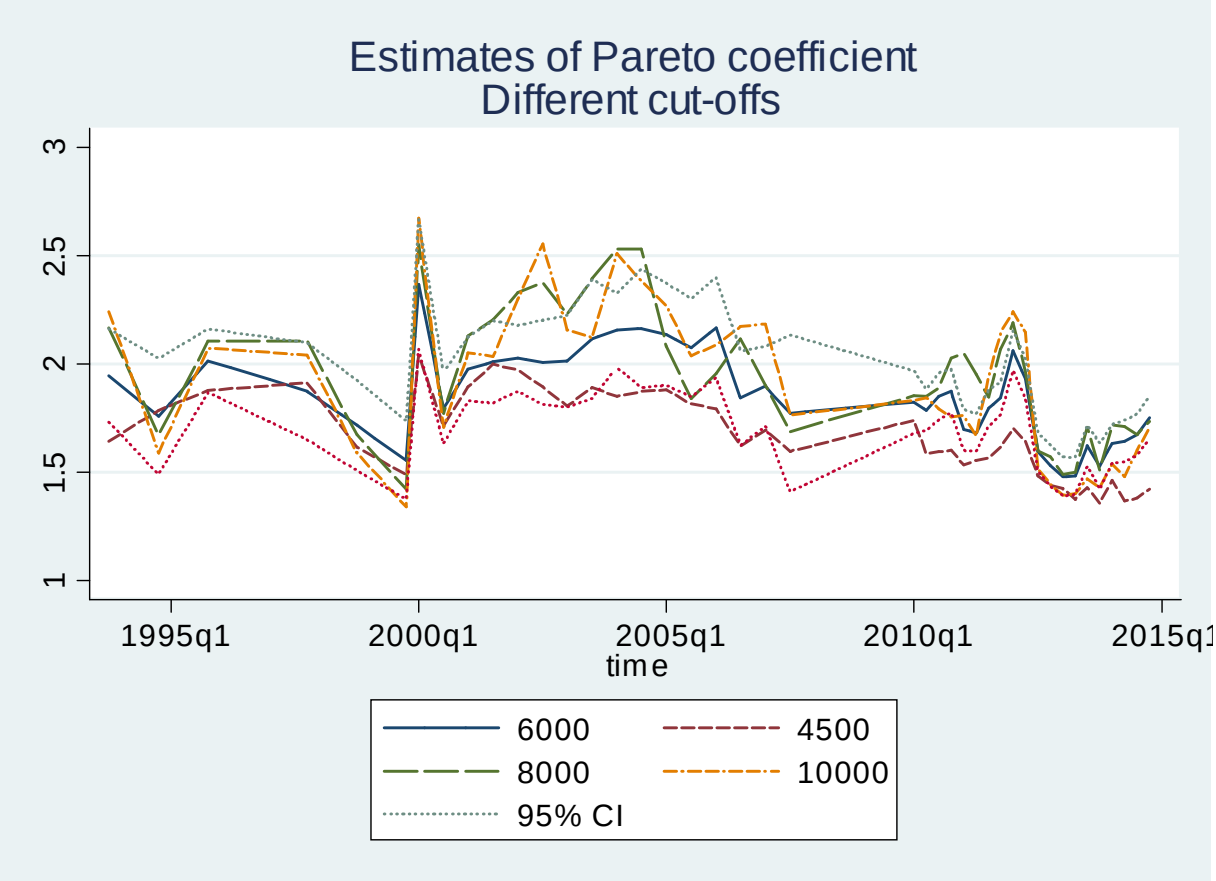


Figure 3: Different cut-offs for the top tail in Rands per month (deflated to June 2000). The 95% confidence band is for the R6000 cut-off.

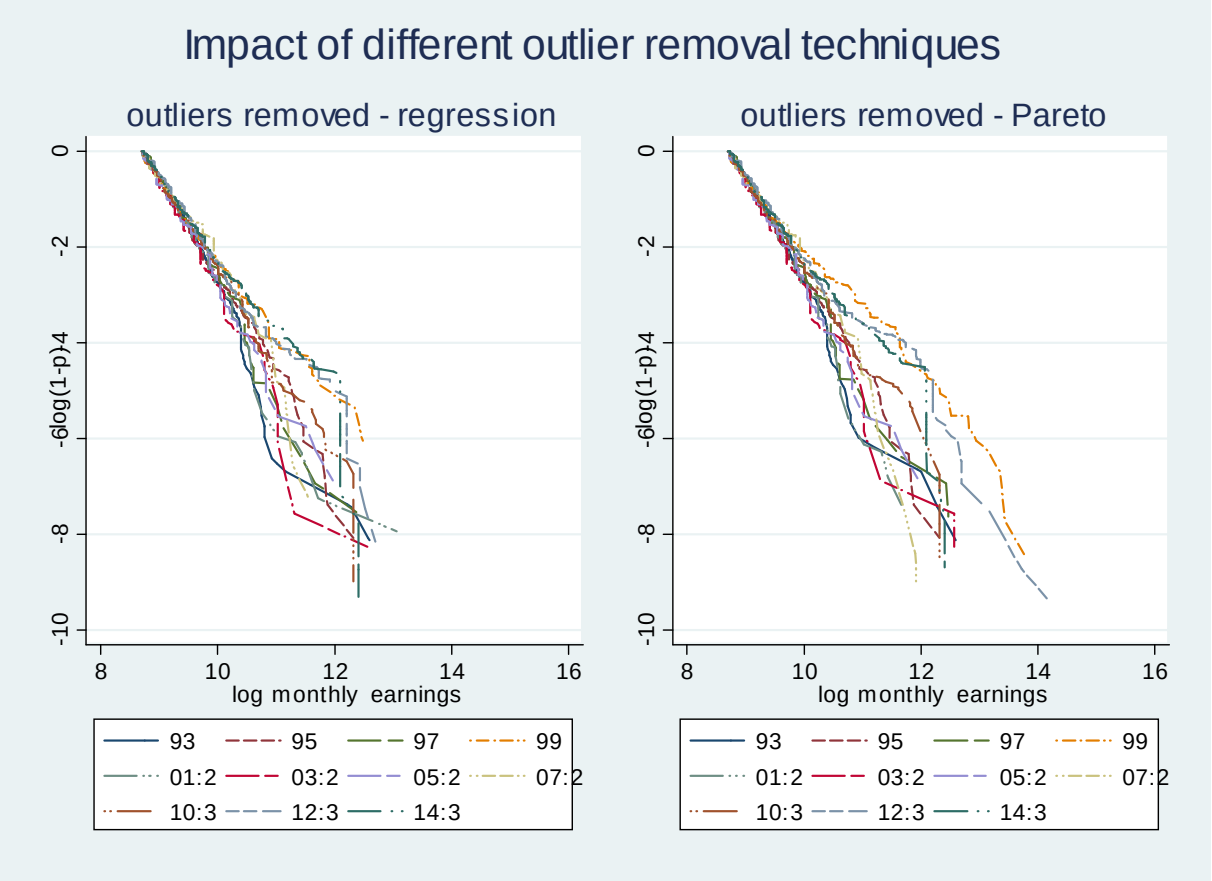


Figure 4: The regression method removes high values much more aggressively than the Pareto procedure outlined in this paper

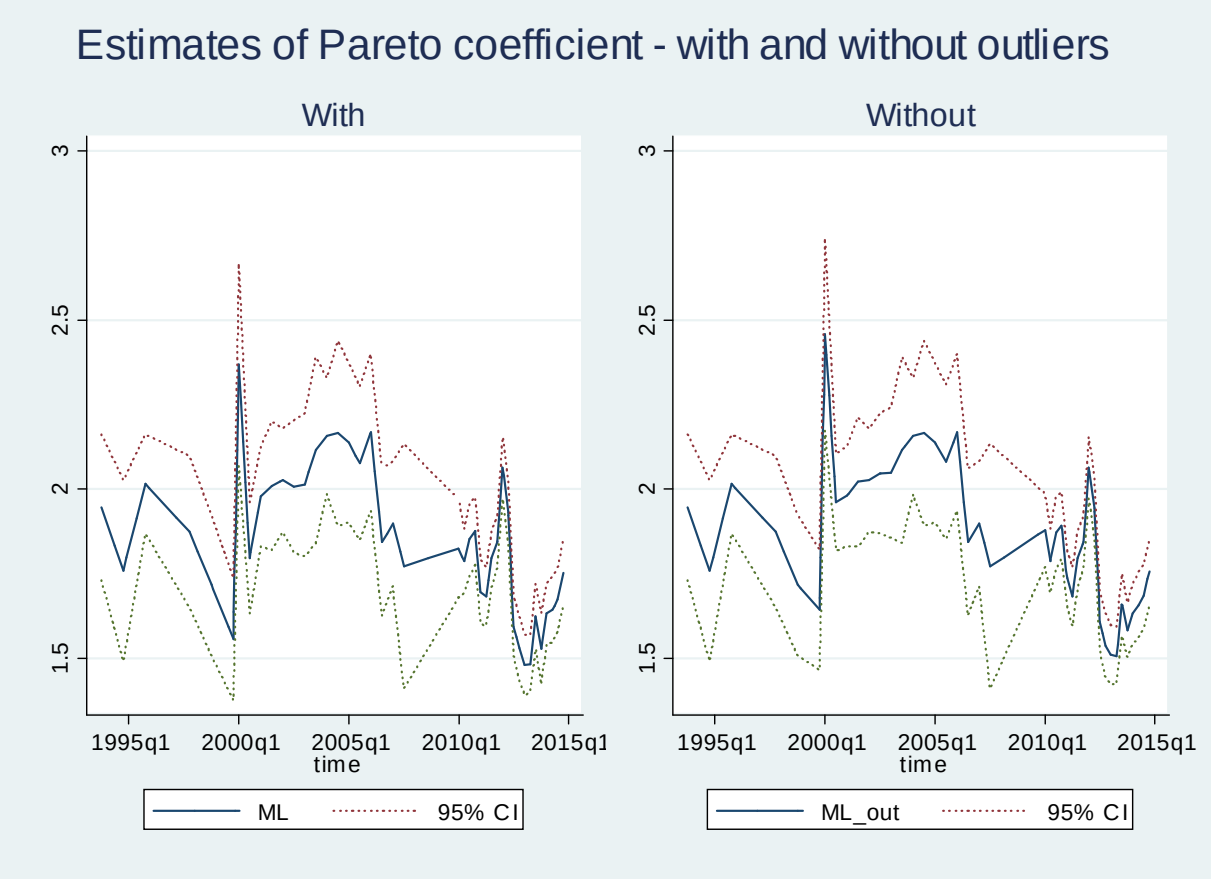


Figure 5: The impact of outlier removal on the estimation of the Pareto coefficient α

5.5 The impact on the estimation of α

At the end of the iterative outlier detection algorithm we obtain an estimate of the Pareto parameter α which is not contaminated by outliers, but as Figure 5 shows the estimates are for the most part hardly affected. This is undoubtedly due to the fact that not that many observations were flagged as outliers through this routine.

5.6 Smoothed Pareto estimation

We noted above that pooling datasets around a particular time "window" (eight quarters) flags even fewer observations as outliers. Nevertheless this procedure has a much stronger impact on the estimation of the Pareto coefficient, as adjoining surveys with quite different tail characteristics (e.g. October 1999 and LFS 2000:1 will be pooled). The impact is shown in Figure 6. The key question is whether pooling the datasets is just smoothing over major measurement shifts that

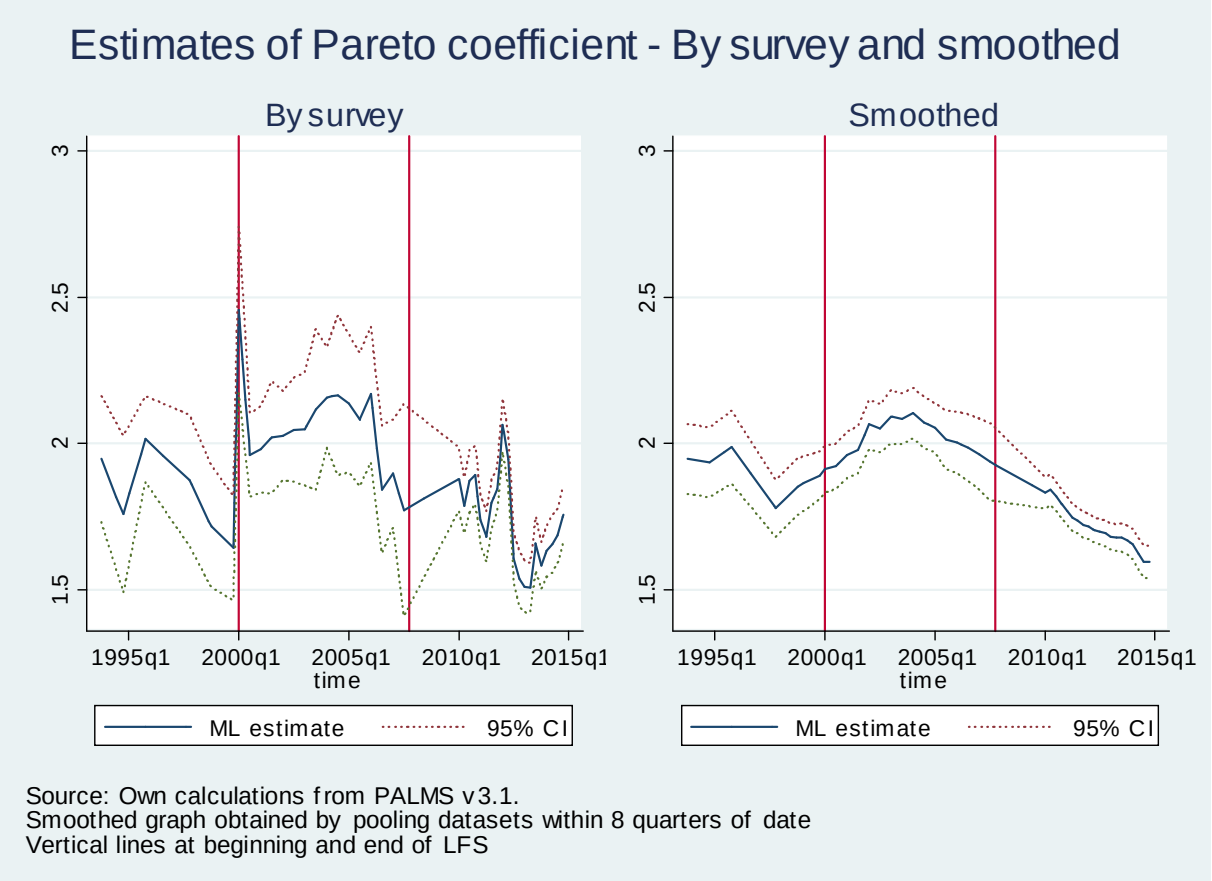


Figure 6: Estimating the tail characteristics by pooling adjoining datasets produces a much smoother time trend of Pareto coefficients.

would otherwise be clearly visible. One thing which the smoothing does make clear is that with the exception of the LFSs (which look anomalous in all sorts of ways) it is clear that the Pareto coefficient is typically below 2 and more likely in the region of 1.8. (Indeed if we were to pool all the available datasets we would get an estimate of 1.79). More provocatively it looks as though the Pareto coefficient is coming down strongly in the recent past, suggesting that the top tail has "thickened".

5.7 Cowell-Flachaire robust estimation of the mean

How big a difference does the semi-parametric technique of Cowell and Flachaire make for the estimation of the mean? A first look at the impact is given in Figure 7 It is evident that the semi-parametric technique does not adequately deal with all types of data contamination. The spikes in October 1999 and September 2000 that exercised Burger and Yu (2007) have not been

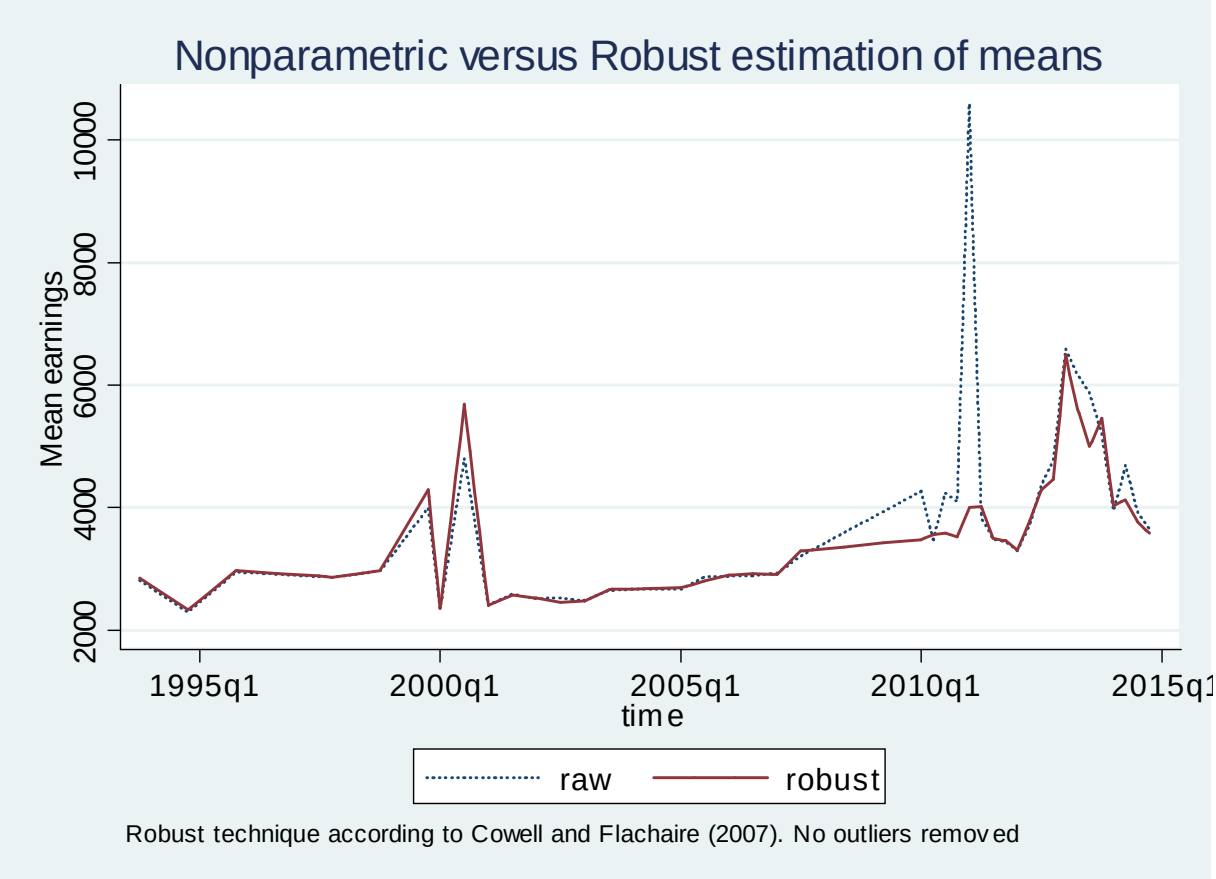
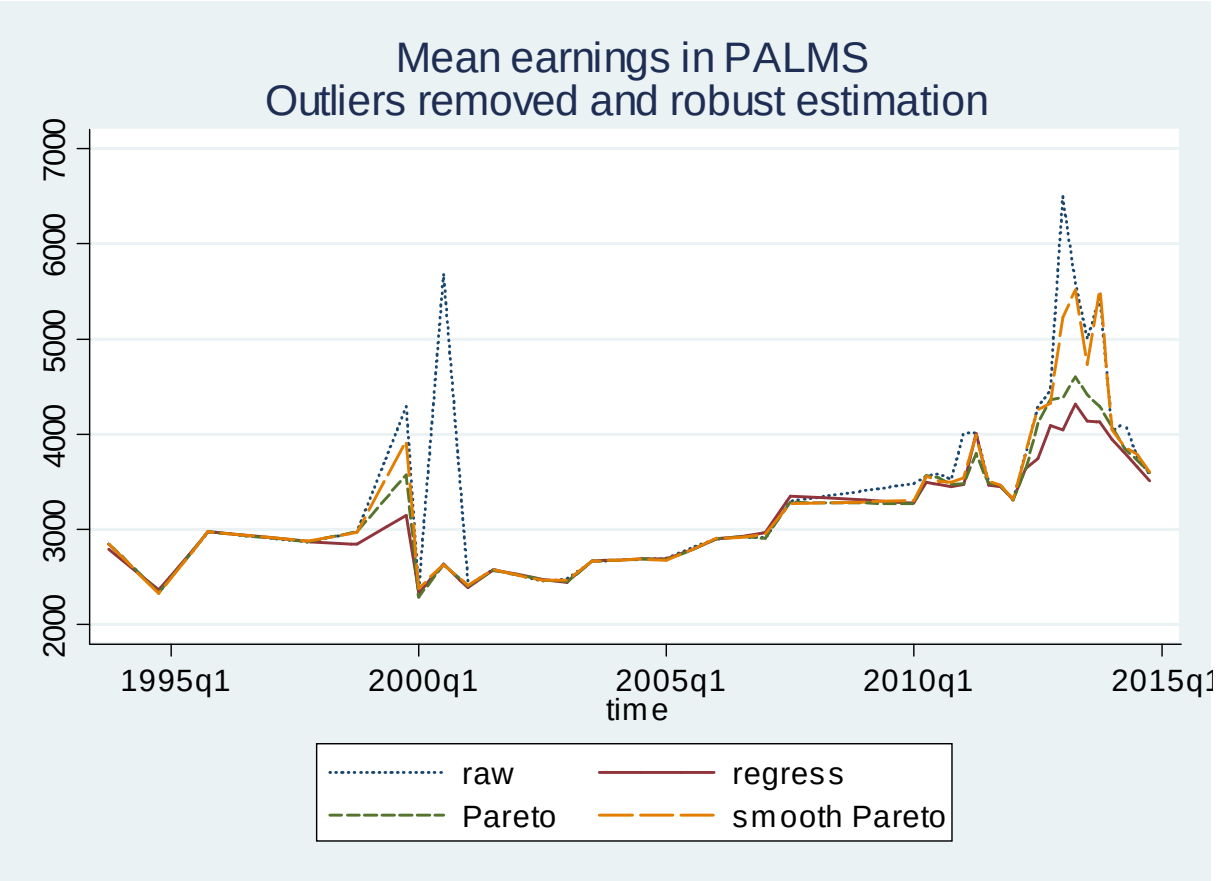


Figure 7: The Semi-parametric technique of Cowell and Flachaire (2007) deals with some data contamination issues, but not all.

removed. Part of the problem may very well be that there are too many contaminating observations. For better or worse we therefore need to combine the technique with some prior outlier detection algorithm. The parametric part of the procedure is therefore geared more at ensuring that the weight of the tail is not underestimated, rather than in removing problematic observations. The impact of the Cowell-Flachaire robust estimation together with different outlier removal routines is shown in Figure 5.7. All of the outlier removal routines get rid of the spike in September 2000, but the October 1999 is removed only with the regression technique. Given what we showed earlier this is not surprising: the problem of OHS 1999 is not one or two outliers, but the entire position of the top tail seems different than in other years. This means several things: first there are many more extreme values to begin with (raising the mean); secondly these don't look anomalous in relation to each other, hence they are less likely to be removed; and thirdly the Pareto coefficient will be particularly low ensuring that the parametric component of the Cowell-Flachaire technique will add weight to the top part of the distribution even if some of the "outliers" are removed.


Mean earnings in PALMS
Outliers removed and robust estimation
2000
3000
4000
5000
6000
7000
1995q1
2000q1
2005q1
2010q1
2015q1
time
raw
regress
Pareto
smooth Pareto

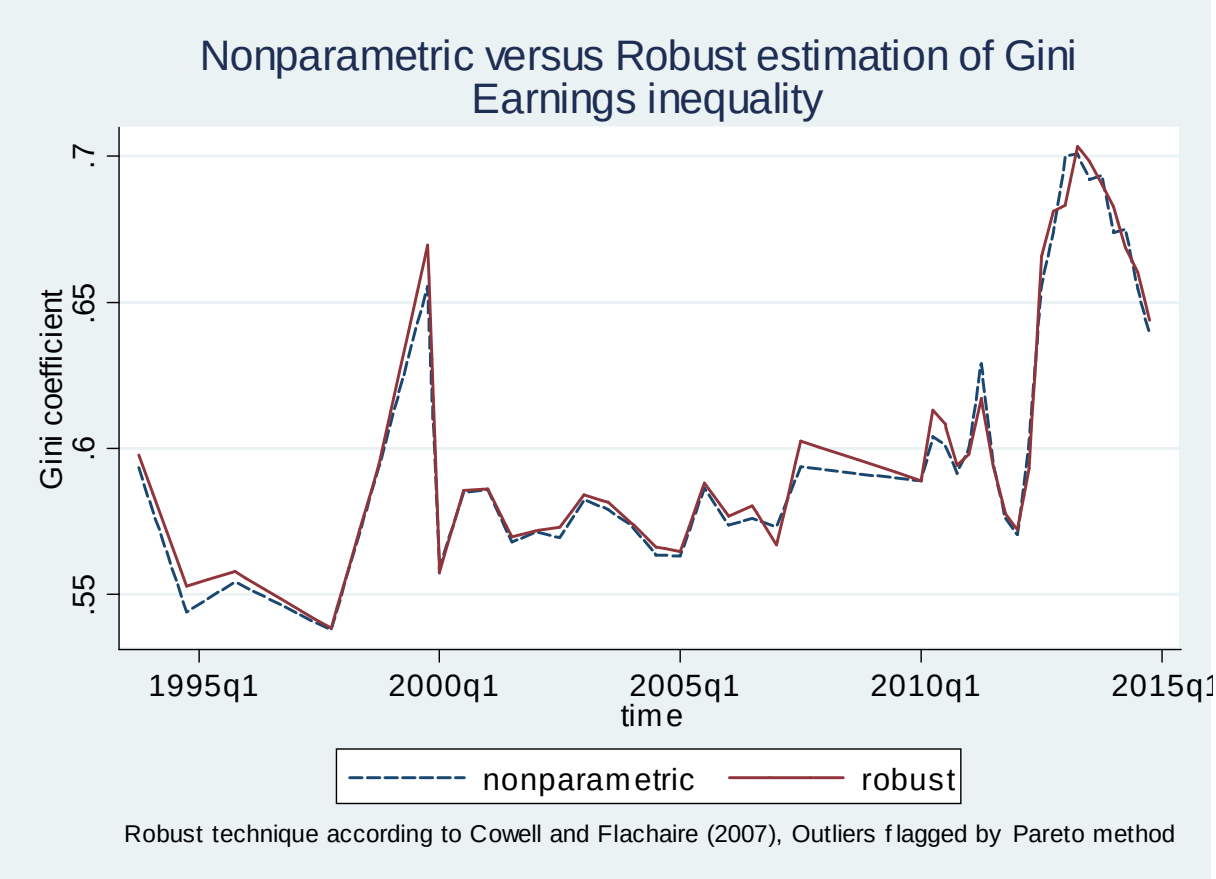


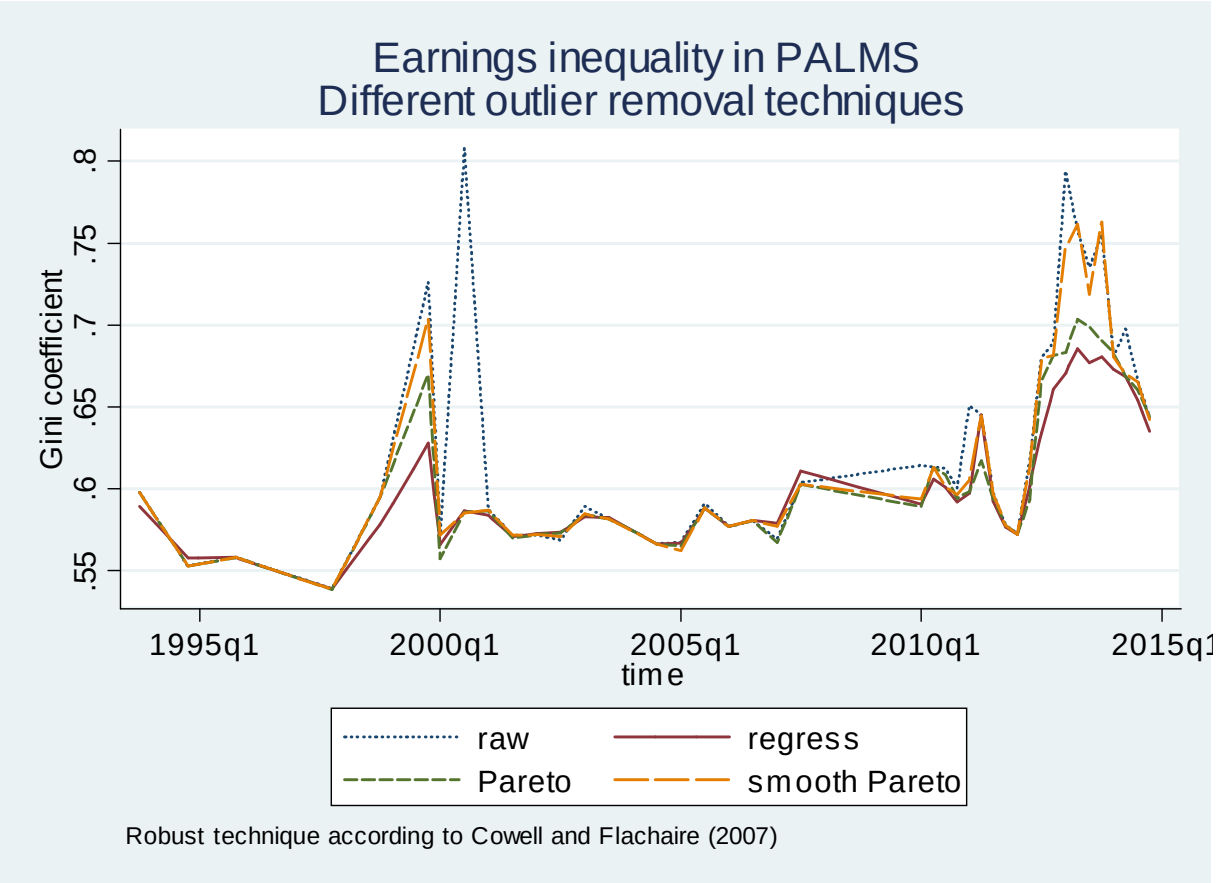
Figure 8: The robust estimation technique of Cowell and Flachaire (2007) tends to increase the Gini coefficient slightly

The OHS 1999 is the most clear-cut case, but we saw in the right panel of Figure 4 that QLFS2012:3 and QLFS2014:3 also had "flatter" profiles. In Figure 5.7 it is also evident that the spikes around 2014 and 2012 are not properly removed by the Pareto based outlier detection programme.

5.8 Cowell-Flachaire robust estimation of the Gini

How does the robust technique fare in relation to the estimation of the Gini? We see that in virtually all cases the robust estimation technique increases the estimated coefficient slightly. But these are small effects – particularly when compared to the big differences that we see between different surveys. Again the huge increase in measured inequality in 1999 and the precipitous drop thereafter are striking.

The impact of the different outlier detection methods is shown in Figure 5.8 which mirrors the



results in relation to mean estimation. It is clear that the spikes in 1999 and 2014 are not due to just one or two problematic data points. The more aggressive outlier removal routine embodied in the regression approach smooths over these bumps, but the underlying issue that needs to be probed is why the top end of the distribution seems so systematically different.

6 Discussion

Our discussion has cycled back to the key issue of measurement. It seems clear that in some of these surveys the "Data Generating Process" was somewhat different from those in other years. Previous discussions have already noted a number of discontinuities between October 1999 and February 2000. The location of the upper end of the earnings distribution can be added to those. Similarly there seem to be differences in some of the waves of the QLFS. One of the big differences between the QLFS data releases and those of the LFSs and the OHSs is that the earnings data is (almost) fully imputed. This leaves plenty of scope for problematic observations to be duplicated. It is therefore at least possible that the 2012 and 2014 QLFS spikes may be due to the imputation

routines. It would, however, be better for the proper academic analysis of the data if it was possible to get access to rawer forms of the data.

Measurement and changes in measurement are first-order effects in trying to understand changes in the earnings distribution over time. Data contamination is clearly another issue. Both the regression and the Pareto routines pick up isolated cases of contamination. Arguably the Pareto approach is better if one wants to understand the nature of the top tail of the distribution. Nevertheless given the fact that measurement shifts seem to swamp these data issues, the more aggressive approach embedded in the regression approach may be more appropriate.

Our estimates suggest that the tails of the income distribution are "heavy" and may even have become heavier over time. This would suggest that the Cowell-Flachaire (2007) procedure should make a difference to our estimates of the mean and of inequality measures. We saw that it made some difference to the estimated Gini coefficient. The impact might be larger on a different measure. Nonetheless it is clear that these impacts will be small relative to the huge shifts induced by other factors. If we can clean up the data properly it may turn out that these techniques may become more relevant.

Of course the fact that our earnings distribution is "heavy tailed" is of interest in its own right. It suggests that the labour market processes are capable of generating considerable inequality. Extreme earnings are more common than the naive reliance on log-normal models would suggest.

7 Conclusion

Our analysis has once again thrown the issue of the measurement into sharp relief. No fancy techniques can undo bad data collection/data entry or processing. But new techniques may help us identify where the problems are and that in turn may help us understand where we have come from and where we are going. In this paper we have developed a new diagnostic tools for outliers which does not blindly remove most extreme values. Instead it tries to identify the sort of extreme values that belong in the top tail and those that don't.

Despite all the "noise" exposed in this paper, there are actually some fairly clear conclusions: the top tail of the earnings distribution is "heavy tailed" with a Pareto coefficient of around 1.8. There is no evidence that this tail is likely to thin out any time soon, in fact the evidence, for what it's worth, is that it is thickening. More substantively, a Pareto coefficient of this magnitude suggests 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.

References

Alvaredo, F. and Atkinson, A. B.: 2010, Colonial rule, apartheid and natural resources: Top incomes in South Africa, 1903-2007, *Discussion Paper 8155*, Centre for Econonomic Policy Research.

Bhorat, H., Van Der Westhuizen, C. and Jacobs, T.: 2009, Income and non-income inequality in post-apartheid South Africa: What are the drivers and possible policy interventions?, *Working Paper 09/138*, DPRU, University of Cape Town. available at http://ssrn.com/abstract=1474271.

Billor, N., Hadi, A. S. and Velleman, P. F.: 2000, BACON: blocked adaptive computationally efficient outlier nominators, *Computational Statistics and Data Analysis* **34**, 279–298.

Blandy, F.: 2009, The rich get richer, Agence France Press, available at http://business.iafrica.com/features/476302.html.

Branson, N. and Wittenberg, M.: 2014, Reweighting South African national household survey data to create a consistent series over time: A cross-entropy estimation approach, *South African Journal of Economics* **82**(1), 19–38.

Burger, R. and Yu, D.: 2007, Wage trends in post-Apartheid South Africa: Constructing an earnings series from household survey data, *Working Paper 07/117*, Development Policy Research Unit, University of Cape Town.

Cowell, F. A. and Flachaire, E.: 2007, Income distribution and inequality measurement: The problem of extreme values, *Journal of Econometrics* **141**, 1044–1072.

Fedderke, J., Manga, J. and Pirouz, F.: 2004, Challenging Cassandra: Household and per capita household income distribution in the October Household Surveys 1995-1999, Income and Expenditure Surveys 1995 & 2000, and the Labour Force Survey 2000, *Working Paper 13*, Economic Research Southern Africa. available at http://www.econrsa.org/system/files/publications/working_papers/wp13.pdf.

Heap, A.: 2009, *Earnings inequality in South Africa: Decomposing changes between 1995 and 2006*, Master's thesis, Economics Department, University of Cape Town.

Hill, B. M.: 1975, A simple general approach to inference about the tail of a distribution, *The Annals of Statistics* **3**(5), 1163–1174.

Kerr, A., Lam, D. and Wittenberg, M.: 2016, Post-Apartheid Labour Market Series [dataset], DataFirst, University of Cape Town. Version 3.1.

Kerr, A. and Wittenberg, M.: 2015, Sampling methodology and field work changes in the October Household Surveys and Labour Force Surveys, *Development Southern Africa* **32**(5), 603–612.

Leibbrandt, M., Finn, A. and Woolard, I.: 2012, Describing and decomposing post-apartheid income inequality in South Africa, *Development Southern Africa* **29**(1), 19–34.

Leibbrandt, M., Woolard, I., Finn, A. and Argent, J.: 2010, Trends in South African income distribution and poverty since the fall of Apartheid, *Social, Employment and Migration Working Papers 101*, OECD. http://dx.doi.org/10.1787/5kmms0t7p1ms-en.

Leite, P. G., McKinley, T. and Osorio, R. G.: 2006, The post-apartheid evolution of earnings inequality in South Africa, 1995-2004, *Working Paper 32*, International Poverty Centre, UNDP.

Mandelbrot, B.: 1960, The Pareto-Lévy law and the distribution of income, *International Economic Review* **1**(2), 79–106.

Neyens, E. and Wittenberg, M.: 2016, Changes in self-employment in the agricultural sector, South Africa: 1994-2012, Working Paper, DataFirst University of Cape Town.

Tregenna, F.: 2011, Earnings inequality and unemployment in South Africa, *International Review of Applied Economics* **25**(5), 585–598.

Tregenna, F. and Tsela, M.: 2012, Inequality in South Africa: The distribution of income, expenditure and earnings, *Development Southern Africa* **29**(1), 35–61.

van der Berg, S.: 2011, Current poverty and income distribution in the context of South African history, *Economic History of Developing Regions* **26**(1), 120–140.

Wittenberg, M.: 2014a, Analysis of employment, real wage, and productivity trends in South Africa since 1994, *Conditions of Work and Employment Series 45*, International Labour Office.

Wittenberg, M.: 2014b, Wages and wage inequality in South Africa 1994-2011: The evidence from household survey data, *Working Paper 135*, Southern Africa Labour and Development Research Unit.