Fitting Pareto Tails to Wealth Survey Data: A Practioners' Guide[1]

Rafael Wildauer
University of Greenwich *

Jakob Kapeller
University Duisburg-Essen and
Johannes Kepler University Linz,
Institute for Comprehensive Analysis of the Economy (ICAE) †

Taking survey data of household wealth as our major example, this short article discusses some of the issues applied researchers are facing when fitting (Type I) Pareto distributions to complex survey data. The contribution of this article is threefold. First, we show how the ordering of the data vector is related to alternative definitions of the empirical $CCDF$. Second, we provide an intuitive reinterpretation of the bias-corrected estimator developed by Gabaix and Ibragimov (2011), in terms of the alternative definitions of the empirical $CCDF$, which allows us to generalize their result to the case of complex survey data. Third, we provide computational formulas for standard Kolmogorov-Smirnov (KS) and Cramer-von Mises (CvM) goodness-of-fit tests for complex survey data. Taken together the article provides a concise and hopefully useful presentation of the fundamentals of Pareto tail-fitting with complex survey data.

Keywords: Pareto distribution, complex survey data, wealth distribution
JEL Classifications: D31, C46, C83

Introduction

Taking survey data on household wealth as our major example, this short article discusses some of the issues applied researchers are facing when fitting (Type I) Pareto distributions to complex survey data. First, we show that Kratz and Resnick's (1996) standard QQ regression approach is often implemented, based on a formulation of the empirical Complementary Cumulative Distribution Function ($CCDF$) that is tailored to avoid the

* *Address for Correspondence:* Department of Economics and International Business, University of Greenwich, Old Royal Naval College, Park Row, Greenwich, London SE10 9LS Email: r.wildauer@gre.ac.uk

† *Address for Correspondence:* Institute for Socio-Economics, University Duisburg-Essen, Forsthausweg 2, 47057 Duisburg, Germany Email: jakob.kapeller@uni-due.de

occurrence of log(0). While often giving plausible results, this approach is not appealing on statistical grounds. In what follows, we provide a different approach to specifying the empirical $CCDF$, which is more general than past approaches and incorporates existing innovations in the literature.

As different formulations of the empirical $CCDF$ typically arise from different orderings of the data vector, we introduce an alternative formulation of the $CCDF$ by averaging the $CCDF$ obtained from data vectors in ascending and descending order. This approach not only allows avoiding the log(0) problem but also corresponds to the bias correction of the QQ regression proposed by Gabaix and Ibragimov (2011). Thus, we are able to provide an alternative, more general and probably also more intuitive explanation of why and how Gabaix and Ibragimov's (2011) bias correction works. At the same time, our alternative formulation is easier to implement in applied work. Finally, with our proposed implementation of Gabaix and Ibragimov (2011), the generalization towards complex survey weights arises naturally. In general the formulation of the averaged $CCDF$ allows for an easy implementation of Vermeulen's (2018) rich list approach or Wildauer and Kapeller's (2019) rank-correction approach.

In addition, we also summarise the extension of goodness-of-fit tests, in the form of the Kolmogorov-Smirnov (KS, from here on) and Cramer-von Mises (CvM, from here on) test statistics, to complex survey data. These extensions are important for researchers interested in determining the scale parameter of the Pareto distribution, not based on ad hoc assumptions or a purely graphical analysis, but on goodness-of-fit tests, as demonstrated by Clauset *et al.* (2009). We close by comparing the QQ regression approach to possible alternatives to underscore the merits of the traditional approach to estimating (Type I) Pareto distributions.

The aim of this article is threefold. First, we provide an alternative perspective on key aspects of Pareto tail-fitting that is easy to implement and technically sound. Second, we show that this approach intuitively captures more intricate issues, such as Gabaix and Ibragimov's (2011) bias correction or the incorporation of complex survey weights. We complement this argument by demonstrating the benefits of this approach of estimating Pareto distributions by means of concrete examples. Third, we summarise key results on goodness-of-fit tests with complex survey data. Overall we think the article provides a concise and useful presentation of the fundamentals of Pareto tail-fitting with complex survey data.

Revisiting Fundamentals

Much of the literature on fitting Pareto tails to wealth survey data (Jayadev 2008, Bach *et al.* 2018, Dalitz 2018, and Vermeulen 2018) relies on simple Ordinary Least Square (OLS) regressions, which fit the shape parameter α of the Pareto distribution to some upper segment of the available data. More specifically, this regression relates the empirical $CCDF$, as derived from the data, to the theoretical $CCDF$, representing the Pareto distribution (Kratz and Resnick 1996). The theoretical $CCDF_T$ for a random variable X following a Pareto distribution is defined as follows:

$$CCDF_T\left(x_i\right) = \Pr\left(X > x_i\right) = \left(\frac{x_m}{x_i}\right)^{\alpha}, \tag{1}$$

where x_m is the scale and α, the shape parameter. In addition, assume we have a random sample of households with net wealth $x = (x_1, \ldots, x_n)$ and corresponding weights $w = (w_1, \ldots, w_n)$, where the number of households represented by the available sample is defined as $N = \sum_{i=1}^{n} w_i$, and the data vector is organised in a descending order (*i.e.*, from the most to the least affluent observation)[2]. This setup yields a data vector, denoted as $x_d = \left(x_{(1)}, \ldots, x_{(n)}\right)$, with the corresponding vector of weights, as $w_d = \left(w_{(1)}, \ldots, w_{(n)}\right)$. Then the empirical $CCDF$ is written as:

$$CCDF\left(x_{(i)}\right)_d = \frac{\sum_{1 \leq j \leq i} w_{(j)}}{N}. \tag{2}$$

In a next step, we assume that some upper segment of the data ($i = (1, .., m)$, where $m < n$) can be described by a Pareto Distribution, which leads to the expression

$$\frac{\sum_{1 \leq j \leq i} w_{(j)}}{N} = \left(\frac{x_m}{x_{(i)}}\right)^{\alpha}. \tag{3}$$

It can be directly translated into the regression equation:

$$\ln\left(\sum_{1 \leq j \leq i} w_{(j)}\right) = C_1 - \alpha \ln\left(x_{(i)}\right) + \epsilon_i, \tag{4}$$

where $C_1 = \ln(N) + \alpha \ln\left(x_m\right)$. Equation 4 is estimated by OLS in order to obtain an estimate of the shape parameter (α). Vermeulen (2018) extends this standard approach along two lines. First, he proposes to add observations from rich lists to the data vector. Second, he incorporates Gabaix and Ibragimov's (2011) (G&I, from here onwards) bias correction and generalizes it to the case of complex survey weights.

For now we will focus on the extension of G&I to complex survey weights. G&I argue that it has long been known that OLS estimation of Equation 4 yields a biased estimate of the shape parameter α (Aigner and Goldberger 1970). They show that subtracting the value 1/2 from an observation's rank, as given by $\sum_{j=1}^{i} w_{(j)}$, will eliminate this bias. G&I also assume, however, that $w_i = w_j = 1$, so that the expression $\sum_{j=1}^{i} w_{(j)}$ is equal to the rank of observation i as represented by the index $(i) = (1), \ldots, (m)$. Vermeulen (2018) extends G&I's bias-correction approach to the case of complex survey weights, which means that weights are not equal to 1 and differ across observations. Starting from a data vector arranged in descending order (x_d) with a corresponding weights vector (w_d), he defines $\bar{N} = \frac{\sum_{i=1}^{m} w_{(i)}}{m}$ as the average weight and $\bar{N}_i = \frac{\sum_{j=1}^{i} w_{(j)}}{i}$ as the average weight up to observation i. Then the empirical $CCDF$ is defined as:

$$CCDF\left(x_{(i)}\right)_d = \frac{\sum_{1 \leq j \leq i} w_{(j)}}{N} = i \frac{\bar{N}_i}{\bar{N}} \frac{\bar{N}}{N}, \tag{5}$$

which allows for a straightforward inclusion of the Gabaix and Ibragimov (2011) bias correction:

$$CCDF\left(x_{(i)}\right)_V = (i - 0.5) \frac{\bar{N}_i}{\bar{N}} \frac{\bar{N}}{N}. \tag{6}$$

Based on that, Vermeulen (2018) derives the equation to be estimated by OLS as:

$$\ln\left((i - 0.5) \frac{\bar{N}_i}{\bar{N}}\right) = C_2 - \alpha \ln\left(x_{(i)}\right) + \epsilon_i, \tag{7}$$

where $C_2 = \ln(N) + \alpha \ln(x_m) - \ln(\bar{N}) = C_1 - \ln(\bar{N})$. This is the standard setting in much of the applied literature to fit a Pareto tail to wealth survey data. Augmenting the regression (Equation 7) by observations from rich lists, as suggested by Vermeulen (2018), is easy to do.

From our definitions above, it follows that the scale parameter x_m represents the minimal value associated with the Pareto distribution. Often this parameter is chosen *ad hoc* by referring to some conventional thresholds or by inspecting visually the relevant data in log-log form. Against this backdrop, it should be mentioned that Clauset *et al.* (2009) suggest a method for determining the scale parameter x_m based on statistical goodness-of-fit tests. The idea is to choose a generous lower bound in the available data, for example, the 90th wealth percentile, and then to fit Pareto distributions to ever smaller sub-samples, created by further increasing the threshold.

In the simplest case one could fit Pareto distributions for increasingly smaller subsets, by successively removing the smallest observation from the data. Then one computes goodness-of-fit statistics based on KS or CvM test statistics to compare the different fitted distributions. The distribution that exhibits the smallest goodness of fit statistic (and thus the best fit to the data) is chosen as the preferred specification. In this way, x_m is defined as the smallest observation in the subset that exhibits the best fit to the data. It is crucial to note that Clauset *et al.*'s (2009) original paper does not discuss the situation of complex survey weights but rather proceeds by suggesting a method for more concisely testing the adequacy of assuming a Pareto tail in the first place.

Against this backdrop, our article comments on three aspects related to these standard procedures. First, we clarify how the ordering of the data impacts the exact shape of the empirical $CCDF$. Second, we provide a simpler and more intuitive formulation of G&I's bias correction from which the generalization to complex survey data emerges naturally. This yields a reformulation of the basic QQ regression incorporating the G&I bias correction and being compatible with Vermeulen's (2018) rich-list approach as well as Wildauer and Kapeller's (2019) rank-correction approach. Third, we discuss the generalization of KS or CvM test statistics to complex survey data.

Rethinking Fundamentals

Data ordering and alternative $CCDF$ definitions

Most exercises in fitting Pareto distributions to wealth survey data start by formatting the data vector in descending order; they then define the empirical $CCDF$ based on Equation 2. This definition does not exactly conform to the theoretical idea expressed in Equation 1, however, which demands a strict inequality and hence denotes the probability of observing someone with wealth greater than x_i. In contrast, the formulation used in Equation 2 does not demand a strict inequality and therefore exhibits a systematic deviation from its theoretical counterpart. In the case of the richest household, $CCDF\left(x_{(1)}\right)_d = \frac{w_{(1)}}{N}$, this implies the assignment of a non-zero probability to observing a household more affluent than $x_{(1)}$, although the theoretical formulation would imply a probability of zero. In the case of the poorest household $CCDF\left(x_{(n)}\right)_d = \frac{N}{N} = 1$, the same inconsistency arises, as a probability of 1 is assigned to this observation instead of a probability slightly less than 1.

The simple reason why the Pareto literature deviates from the theoretical definition of the $CCDF$ and defines the latter based on Equation 2 is that it ensures that the empirical $CCDF$ is never equal to 0 and can thus be log-linearised without having to deal with $\log(0) = -\infty$. This is the reason why other applications, which do not require log-linearization, organise the data vector in ascending order $x_a = \left(x_{(n)}, \ldots, x_{(1)}\right)$ with a corresponding vector of weights $w_a = \left(w_{(n)}, \ldots, w_{(1)}\right) = (1, \ldots, 1)$ and then define the empirical $CCDF$ as:

$$CCDF\left(x_{(i)}\right)_a = 1 - \frac{\sum_{i \leq j \leq n} w_{(j)}}{N}. \tag{8}$$

This approach to the empirical $CCDF$ is perfectly in line with its theoretical counterpart. For the case of the richest household, $CCDF\left(x_{(1)}\right)_a = 1 - \frac{N}{N} = 0$, a zero probability is assigned to observing a more affluent household, and for the poorest household, $CCDF\left(x_{(n)}\right)_a = 1 - \frac{w_{(n)}}{N} = 1 - \frac{1}{N}$, a below unit probability is assigned to observe a more affluent household. These examples demonstrate why a definition based on Equation 8 more accurately represents the available sample information. The downside is that if log linearization is required as when using OLS to fit a Type I Pareto tail, the richest observation with $CCDF\left(x_{(1)}\right)_a = 1 - \frac{N}{N} = 0$ would need to be discarded, and so consistency with the theoretical $CCDF$ comes at the cost of losing one data point in a rather non-random way.

That said, we can derive formulas to compute $CCDF\left(x_{(i)}\right)_a$ from data organised in descending order and vice versa. For example, given the data vector $x_d = \left(x_{(1)}, \ldots, x_{(n)}\right)$ with corresponding weights $w_d = \left(w_{(1)}, \ldots, w_{(n)}\right)$, we can define $w_{(0)} = 0$. Thus, we amend the weights vector with a zero entry and then compute $CCDF\left(x_{(i)}\right)_a$, as:

$$CCDF\left(x_{(i)}\right)_a = \frac{\sum_{1 \leq j \leq i} w_{(j-1)}}{N}. \tag{9}$$

Note that the main difference between Equation 8 and Equation 9 are the two different subsets of observations over which the sum operator is defined, $i \leq j \leq n$ in the first case and $1 \leq j \leq i$ in the latter. If both equations are applied on the same weights vector, they will give identical results.

So the issue of defining the empirical $CCDF$ is not about how the data is organised, but from a conceptual point of view, whether one wants to define the empirical $CCDF$ as equal to the probability $\Pr\left(X > x_{(i)}\right)$, as in the case of CCDF $\left(x_{(i)}\right)_a$ or equal to the probability $\Pr\left(X \geq x_{(i)}\right)$, as in the case of CCDF $\left(x_{(i)}\right)_d$. The reason why the literature on fitting Pareto tails chooses the definition $CCDF\left(x_{(i)}\right)_d = \Pr\left(X \geq x_{(i)}\right)$ by default simply lies in the need to log-linearise the $CCDF$ and to avoid the logarithm of 0.

In order to assess the impact of these different definitions on the estimation of Pareto tails we conducted a Monte Carlo simulation of the following form. We drew a sample of $n = 500$ from a Type I Pareto distribution with $P(\alpha, x_m)$. We then used four different estimators to estimate α: three OLS estimators using definitions of the empirical $CCDF$ based i) on ascending (OLS_a), ii) on descending (OLS_d) ordering, and iii) on the average of these two orderings (OLS_{av}); iv) we also applied, as a point of reference, a standard maximum likelihood estimator (MLE).[3] We drew 10,000 samples for a range of empirically plausible shape parameters $(1, 1.2, 1.4, 1.6, 1.8, 2)$ and fixed the scale parameter at $x_m = 10^6$. The results are summarised in Table 1.

The OLS estimator based on a descending ordering (OLS_d) systematically underestimates the true shape parameter depicted in the first column and thus overestimates the thickness of the tail. This is rather unsurprising, given that $CCDF\left(x_{(i)}\right)_d$ is larger than the theoretically accurate expression $CCDF\left(x_{(i)}\right)_a$ and thus implies a thicker tail. While the OLS estimator based on an ascending ordering (OLS_a) is based on $CCDF\left(x_{(i)}\right)_a$, it ignores the richest observation and therefore systematically overestimates the true shape parameter and underestimates the thickness of the tail. The OLS estimator that relies on the average of the $CCDF$ of the other two OLS estimators (OLS_{av}) yields on average a shape parameter that is substantially closer to the true underlying parameter and yields an even smaller absolute deviation from the latter than the maximum likelihood estimator (MLE). We will show in the next section that OLS_{av} is equivalent to the G&I bias correction. Before deriving this equivalence result, Table 2 sheds some light on the economic significance of the seemingly minor differences in the shape parameter reported in Table 1.

Table 2 shows the deviation of the estimated tail wealth (based on estimated alphas) relative to actual tail wealth according to the true underlying Pareto distribution. This means that for an underlying Pareto distribution with $\alpha = 1.2$, the OLS_d estimator yields a tail wealth estimate of 111.2% of the true amount, while OLS_a yields a total tail wealth estimate of 93.3% of the true underlying total. OLS_{av} yields a fairly accurate estimate of 101.2% of the underlying total, and MLE yields an estimate of 98.7% and thus performs slightly worse than OLS_{av} in absolute terms. The bias across all estimators is more severe for highly concentrated populations with small Pareto alphas.

Overall, the Monte Carlo study shows that the OLS estimator that relies on the averaged CCDF definition (OLS_{av}) outperforms both OLS_a and OLS_d and even slightly outperforms MLE. This result can be rationalized

Table 1
Estimated shape parameters (Pareto alpha)

alpha		OLS_d	OLS_{av}	MLE	OLS_a
1.0	mean	0.978	0.998	1.003	1.013
	std	0.06	0.06	0.05	0.06
1.2	mean	1.173	1.197	1.203	1.216
	std	0.07	0.08	0.05	0.07
1.4	mean	1.369	1.397	1.405	1.419
	std	0.09	0.09	0.06	0.08
1.6	mean	1.565	1.597	1.606	1.622
	std	0.10	0.10	0.07	0.10
1.8	mean	1.759	1.795	1.806	1.824
	std	0.11	0.11	0.08	0.11
2.0	mean	1.955	1.995	2.007	2.026
	std	0.12	0.12	0.09	0.12

Reported values represent the mean and standard deviation of the estimated Pareto alpha based on samples of size $n = 500$ from Pareto distributions with alphas between 1 and 2 and $x_m = 10^6$. Each cell is based on 10,000 independent draws.

by interpreting OLS_{av} as a compromise between theoretical rigor (OLS_a) and the use of all available sample information (OLS_d) when specifying the empirical $CCDF$. The next section will demonstrate that OLS_{av} is equivalent to the bias correction proposed by G&I.

Bias Correction by Averaging

G&I argue that in order to reduce the bias in OLS based estimates of the shape parameter of the Pareto distribution, researchers should subtract the value 1/2 from the rank of each observation prior to performing the rank-wealth regression. The rank of observation i in their terminology is equivalent to the cumulative weight of that observation i given by $\sum_{j=1}^{i} w_{(j)}$ based on a descending data vector. So the bias correction proposed by Gabaix and Ibragimov (2011) corresponds to computing the empirical $CCDF$ in the following way:

$$\text{CCDF}\left(x_{(i)}\right)_{G\&I} = \frac{\left(\sum_{1\leq j\leq i} w_{(j)}\right) - 0.5w_{(i)}}{N}. \tag{10}$$

We can reformulate $CCDF\left(x_{(i)}\right)_{G\&I}$ in the following way after defining $w_{(0)} = 0$:

Table 2
Estimated tail wealth in % of actual tail wealth

alpha	OLS_d	OLS_{av}	MLE	OLS_a
1.2	111.2%	101.2%	98.7%	93.3%
1.4	105.5%	100.5%	99.1%	96.6%
1.6	103.6%	100.3%	99.4%	97.7%
1.8	102.8%	100.3%	99.6%	98.3%
2.0	102.2%	100.2%	99.6%	98.7%

Results based on results from Monte Carlo simulation reported in Table 1.

$$\text{CCDF}\left(x_{(i)}\right)_{AV} = \frac{\left(\sum_{1\leq j\leq i} w_{(j)}\right) - 0.5 w_{(i)}}{N} = \frac{2\left(\sum_{1\leq j\leq i} w_{(j)}\right) - w_{(i)}}{2N} = \frac{\sum_{1\leq j\leq i} w_{(j-1)} + \sum_{1\leq j\leq i} w_{(j)}}{2N} = \left[\text{CCDF}\left(x_{(i)}\right)_a + CCDF\left(x_{(i)}\right)_d\right]/2. \tag{11}$$

It is now clear that Equation 11 is nothing more than the average between the two previous definitions of the empirical $CCDF$. Thus the bias correction proposed by Gabaix and Ibragimov (2011) is equivalent to OLS_{av} in the Monte Carlo simulation in the previous section. Demonstrating this equivalence is useful for two reasons. First, OLS_{av} provides a simple intuition for why the G&I bias correction works and is necessary. It does not rely on the same lengthy and rather technical derivation. Second, with Equation 11 it is straightforward to generalize this result to the case of complex survey weights, as the formula takes weights explicitly into account. Hence, it does not matter whether all weights are implicitly assumed to be equal to 1 $w = (w_1, \ldots, w_n) = (1, \ldots, 1)$ or whether they are allowed to be different for each observation: $w = (w_1, \ldots, w_n)$ and $w_j \neq w_i$.

This last point becomes clear when looking at the way in which Vermeulen (2018) incorporates complex survey weights into his OLS based estimator. The attempt to preserve the individual rank i for each observation in order to subtract 0.5 as a way of implementing the G&I bias correction yields an overly complicated expression, as outlined in Equation 7. In general, $\text{CCDF}\left(x_{(i)}\right)_{AV}$ can be used as the basis for fitting a Pareto distribution in every practically relevant case: for analysing a given sample of (complex) survey data as it is or as a starting point for applying either Vermeulen's (2018) rich-list approach or Wildauer and Kapeller's (2019) rank-correction approach. Effectively it suggests a versatile and intuitive way to implement the G&I bias correction in applied work.

In the remaining section we will explore the performance of the four estimators from the previous section and additionally consider Vermeulen's (2018) estimator (OLS $_{VER}$), which implements the G&I bias correction (Equation 7). We will use these five estimators to fit a Type I Pareto distribution to data from the Survey of Consumer Finances (SCF), 2016. The latter is widely regarded as one of the most reliable surveys on household wealth due to its effective oversampling of affluent households based on income tax information, which is employed in the sample design. Table 3 depicts the estimated shape parameters using all five estimators and applies them to SCF data, while using different quantiles as the cut-off for defining the scale parameter x_m. The largest cut-off restricts the sample used to the most affluent 0.05% of the population (Column 2) and the lowest cut-off limits the sample to the households belonging to the most affluent 4% of the population of the United States (last column of Table 3).

The results are broadly in line with the results from the Monte Carlo simulation using synthetic data. OLS_a consistently yields the highest estimates of the shape parameter and thus is likely to underestimate the tail thickness. OLS_d produces consistently lower estimates and OLS_{av} and OLS_{VER} are located in between these two results as expected from the Monte Carlo simulation. The difference between the latter two vanishes with larger cut-offs and is not relevant in practice[4]. The key difference to the Monte Carlo simulation is the performance of the maximum likelihood estimator.

This latter result indicates that the SCF data is still suffering from some form of differential nonresponse bias. In this case, Vermeulen (2018) showed that the MLE estimator produces lower estimates of the shape parameter than OLS_{VER}. The fact that the SCF explicitly excludes extremely wealthy individuals who show up on the Forbes 400 list can be interpreted as a special case of differential nonresponse. Hence, the richest 400 Americans are excluded by design. In Wildauer and Kapeller (2019), we show that this problem can be dealt with by means of a simple correction and refer the interested reader to this article [5].

Goodness-of-fit statistics and complex survey weights

When fitting Pareto tails to available data, the obvious follow-up question is to ask whether the estimated distribution represents a good fit. In addition, determining the scale parameter of the Pareto distribution by relying on Clauset *et al.*'s (2009) method instead of the common approach of inspecting log-log plots, requires a goodness-of-fit test. In both instances, goodness-of-fit tests need to be adapted to deal with complex survey data. These results have already been established in the literature. Restating them here serves

Table 3
Estimated Pareto alphas for the Survey of Consumer Finances 2016

	0.05 %	0.01 %	0.1 %	0.5 %	1 %	1.5 %	2 %	2.5 %	3 %	4 %
MLE	2.066	1.969	1.739	1.577	1.590	1.512	1.336	1.229	1.151	1.119
OLS_d	2.114	2.000	1.646	1.679	1.653	1.644	1.622	1.595	1.567	1.517
OLS_{av}	2.177	2.041	1.659	1.686	1.659	1.649	1.627	1.600	1.571	1.520
OLS_{VER}	2.191	2.052	1.662	1.688	1.660	1.650	1.628	1.600	1.571	1.521
OLS_a	2.370	2.161	1.691	1.703	1.672	1.661	1.637	1.609	1.579	1.527

Note: Each row presents results for one of the estimators of interest, and the columns depict the chosen cut-off as the richest x% of American households.

the purpose of providing a concise summary and guide for practitioners, who might not be aware of the specialised statistical literature (Monahan 2011, D'Agostino & Stephens 1986.)

We will begin with the KS test. The test statistic (T_{KS}) for unweighted data or trivial weights $w = (1, \ldots, 1)$ is defined as follows (Monahan 2011, p. 351):

$$T_{KS} = \sqrt{n}D, \tag{12}$$

where D is the KS statistic and is defined as:

$$D = \sup_y |CDF_E(x) - CDF_T(x)| \tag{13}$$

which is the maximum distance between the empirical distribution function CDF_E and the theoretical distribution function CDF_T. With a data vector in descending order $x_d = \left(x_{(1)}, \ldots, x_{(n)}\right)$ and a corresponding vector of weights $w_d = (1, \ldots, 1)$ and $w_0 = 0$, the empirical CDF is defined as:

$$CDF\left(x_{(i)}\right)_a = 1 - \frac{\sum_{1 \leq j \leq i} w_{(j-1)}}{N}. \tag{14}$$

For trivial weights, the KS statistic can be computed as follows (Monahan 2011, p. 351):

$$D = \max_i \left(\frac{i}{n} - CDF_T\left(x_{(i)}\right), CDF_T\left(x_{(i)}\right) - \frac{i-1}{n} \right) \tag{15}$$

where n represents the number of observations in the sample. For nontrivial weights $w_d = \left(w_{(1)}, \ldots, w_{(n)}\right)$, the KS statistic is defined as follows (Monahan 2011, p. 358):

$$D = \max_i \left(CDF\left(x_{(i)}\right)_a - CDF_T\left(x_{(i)}\right), CDF_T\left(x_{(i)}\right) - CDF\left(x_{(i-1)}\right)_a \right). \tag{16}$$

The test statistic for nontrivial weights can then be obtained as (Monahan 2011, p. 358):

$$T_{KS} = D\sqrt{\frac{N^2}{\sum_{i=1}^{n} w_{(i)}^2}}. \tag{17}$$

For the CvM goodness-of-fit test, the test statistic (T_{CvM}) for trivial weights is defined as follows (D'Agostino and Stephens 1986, p. 101):

$$T_{CvM} = nW^2 = \frac{1}{12n} + \sum_{i=1}^{n}\left(CDF_T\left(x_{(i)}\right) - \frac{i-0.5}{n}\right)^2. \tag{18}$$

Deriving this expression relies on the probability integral transformation. The CvM criterion $\left(W^2\right)$ is defined as:

$$W^2 = \int_{-\infty}^{\infty}\left[CDF_E - CDF_T\right]^2 dCDF_T. \tag{19}$$

We define $CDF_T\left(x_{(i)}\right) = U_i$, and the probability integral transformation allows us to rewrite the definition of the CvM test statistic as:

$$W^2 = \int_{-\infty}^{\infty}\left[CDF_E - U\right]^2 dU, \tag{20}$$

which simplifies to Equation 18 after multiplying by n.

Deriving the CvM test statistic for the case of complex survey weights is a bit more complicated. Starting from Equation 20, we have to split up the integral:

$$W^2 = \int_{-\infty}^{x_{(1)}}[0-U]^2 dU + \sum_{i=1}^{n-1}\int_{x_{(i)}}^{x_{(i+1)}}\left[CDF\left(x_{(i)}\right)_E - U\right]^2 dU + \int_{x_{(n)}}^{\infty}[1-U]^2 dU. \tag{21}$$

After evaluating the integrals we obtain:

$$\mathrm{W}^2 = \tfrac{1}{3}U_1^3 + \tfrac{1}{3}\sum_{i=1}^{n-1}\left[\left(CDF\left(x_{(i)}\right)_a - U_i\right)^3 - \left(CDF\left(x_{(i)}\right)_a - U_{i+1}\right)^3\right] + \tfrac{1}{3}\left[1-U_n\right]^3. \tag{22}$$

The test statistic is then obtained as

$$T_{CvM} = nW^2. \tag{23}$$

Conclusion

In this article, we make three contributions. First, we show how the ordering of the data vector is related to alternative definitions of the empirical $CCDF$. The different estimators for the shape parameter of the Type I Pareto distribution, which emerge from the competing definitions of the $CCDF$, are compared in a Monte Carlo simulation. The key result is that the estimator based on the average between the $CCDF$ obtained from a data vector in ascending and descending order yields close to unbiased shape parameter estimates. Second, we show that this first result is an alternative and presumably faster and easier derivation of the bias-corrected estimator developed in Gabaix and Ibragimov (2011). This is not only useful because of the straightforward intuition our results provide but also because the average formulation can be directly generalized to the case of household survey data with complex survey weights. The latter case is especially important for researchers working on the distribution of household income and wealth. Third, we provide computational formulas for standard KS and CvM goodness-of-fit tests for complex survey data. These are required in order to determine the length of the Pareto tail systemically rather than by relying solely on graphical methods or conventional cut-off points.

Appendix

Starting from a net wealth vector $x_d = \left(x_{(1)}, \ldots, x_{(n)}\right)$ in descending order (and the corresponding vector of weights $w_d = \left(w_{(1)}, \ldots, w_{(n)}\right)$), the Monte Carlo simulation and the application to Survey of Consumer Finances data are based on the following estimation approaches. The OLS_d estimator of the shape parameter (α) is obtained from estimating the following regression by OLS:

$$\ln\left(\sum_{j=1}^{i} w_{(j)}\right) = C_1 - \alpha \ln\left(x_{(i)}\right) + \epsilon_i, \tag{A1}$$

where $C_1 = \ln(N) + \alpha \ln(x_m)$. The OLS_a estimator is obtained from the following regression:

$$\ln\left(\sum_{j=0}^{i-1} w_{(j)}\right) = C_1 - \alpha \ln\left(x_{(i)}\right) + \epsilon_i, \tag{A2}$$

where the most affluent observation $(i.e.,\ x_{(1)})$ is dropped from the data in order to avoid ln(0). The OLS_{VER} estimator is obtained from the following regression:

$$\ln\left((i-0.5)\bar{N}_i\right) = C_1 - \alpha \ln\left(x_{(i)}\right) + \epsilon_i, \tag{A3}$$

where $\bar{N}_i = \left(\sum_{j=1}^{i} w_{(j)}\right)/i$. The OLS_{av} estimator is obtained from the following regression:

$$\ln\left(\left[\sum_{j=1}^{i} w_{(j)} + \sum_{j=0}^{i-1} w_{(j)}\right]/2\right) = C_1 - \alpha \ln\left(x_{(i)}\right) + \epsilon_i, \tag{A4}$$

where $w_0 = 0$.

Notes

[1] *Acknowledgements:* We thank an anonymous referee for helpful comments as well as Stefan Steinerberger for continued support. All remaining errors are ours.

[2] We thank an anonymous referee to have pointed out that i is simply the index across all observations and does not refer to the complex number $\sqrt{-1}$. We follow the literature and use the term "complex survey weights" to refer to heterogeneous weights, which arise from sampling procedures, such as stratification.

[3] See the Appendix for the exact definitions of these estimators.

[4] The difference between OLS_{av} and OLS_{VER} arises due to the fact that the latter calculates $CCDF\,(x_i)$ based on a correction of the average over the weights observed up to observation i (see Equation 6). In contrast, OLS_{av} calculates the empirical $CCDF\,(x_i)$ which is based on a correction of the weight of observation i only (see Equation 10). So whenever $w_{(i)}$ is very different from the average weight up to observation i (*i.e.* $\bar{N}_i = \left(\sum_{j=1}^{i} w_{(j)}\right)/i\Big)$, the two ways of calculating the empirical $CCDF$ will be different.

[5] Since the SCF suffers from such a differential nonresponse problem, which is not the focus of this article, and OLS_a effectively partly corrects for this problem by dropping the most affluent observation, we dropped the most affluent observation from all specifications. For a detailed discussion of the differential nonresponse problem which arises from excluding the Forbes 400 in the SCF, see Wildauer and Kapeller 2019.

References

Aigner, D.J. and A.S. Goldberger 1970 "Estimation of Pareto's law from grouped observation", *Journal of the American Statistical Association* 65(330): 712-723.

Bach, S., A. Thiemann, and A. Zucco 2018 "Looking for the missing rich: Tracing the top tail of the wealth distribution", DIW Berlin Discussion Paper.

Clauset, A., C.R. Shalizi and M.E.J. Newman 2009 "Power-law distributions in empirical data", *SIAM Review* 51(4): 661-703.

D'Agostino, R.B. and M.A. Stephens (eds) 1986 Goodness-of-Fit Techniques *in D.B. Owen (ed.), Statistics: Textbooks and Monographs* Vol. 68. New York: Marcel Dekker, Inc.

D'Agostino, R. B. and M.A. Stephens 1986 "Goodness-of-fit-techniques" in D.B. Owen (Ed.), *Statistics: Textbooks and Monographs* Vol. 68. New York: Marcel Dekker Inc.

Dalitz, C. 2018 "Estimating wealth distribution: Top tail and inequality", Cornell University arXiv, preprint arXiv 1807.03592.

Gabaix, X. and R. Ibragimov 2011 "Rank - 1/2: A simple way to improve the OLS estimation of tail exponent" *Journal of Business and Economic Statistics* 29(1): 24-39.

Jayadev, A. 2008 "A power law tail in India's wealth distribution: Evidence from survey data", *Physica A: Statistical Mechanics and its Applications* 387(1): 270-276.

Kratz, M. and S.I. Resnick 1996 "The qq-estimator and heavy tail", *Communications in Statistics. Stochastic Models* 12(4): 699-724.

Monahan, J. F. 2011 *Numerical Methods of Statistics* 2nd ed. New York: Cambridge University Press.

Survey of Consumer Finances 2016. The study is sponsored by the Federal Reserve Board in cooperation with the Department of the Treasury. Data have been collected by the National Opinion Research Center (NORC) at the University of Chicago.

Vermeulen, P. 2018 "How fat is the top tail of the wealth distribution?", *Review of Income and Wealth* 64(2): 357-387.

Wildauer, R. and J. Kapeller 2019 "Rank correction: A new approach to differential nonresponse in wealth survey data", *Greenwich Papers in Political Economy* 73.