Policy Research Working Paper 10231

Outlier Detection for Welfare Analysis

Federico Belotti
Giulia Mancini
Giovanni Vecchi

WORLD BANK GROUP
Poverty and Equity Global Practice
November 2022

Policy Research Working Paper 10231

Abstract

Extreme values are common in survey data and represent a recurring threat to the reliability of both poverty and inequality estimates. The adoption of a consistent criterion for outlier detection is useful in many practical applications, particularly when international and intertemporal comparisons are involved. This paper discusses a simple, univariate detection procedure to flag outliers in the distribution of any variable of interest. It presents outdetect, a Stata command that implements the procedure and provides useful diagnostic tools. The output of outdetect compares statistics—with focus on inequality and poverty measures—obtained before and after the exclusion of outliers. Finally, the paper carries out an extensive sensitivity exercise, where the same outlier detection method is applied consistently to per capita expenditure across more than 30 household budget surveys. The results are clear-cut and provide a sense of the influence of extreme values on poverty and inequality estimates.

This paper is a product of the Poverty and Equity Global Practice. It is part of a larger effort by the World Bank to provide open access to its research and make a contribution to development policy discussions around the world. Policy Research Working Papers are also posted on the Web at http://www.worldbank.org/prwp. The authors may be contacted at gvecchi@worldbank.org.

Outlier Detection for Welfare Analysis

Federico Belotti, Giulia Mancini, and Giovanni Vecchi[1]

Keywords: outliers, extreme values, inequality, poverty, incremental trimming curve.

JEL codes: I32, C87, D60.

[1] Federico Belotti and Giovanni Vecchi are Professors at the University of Rome "Tor Vergata", Department of Economics and Finance; Giulia Mancini is a Research Fellow at the University of Sassari, Italy, Department of Economics and Business. All three authors are consultants to the World Bank. This paper was prepared under the supervision of the Poverty & Equity Global Practice Global Solutions Group 1: Data for Policy, and benefitted from the oversight of Utz Pape (World Bank). We thank Silvia Redaelli and Alexandra Jarotschkin for useful comments shared while reviewing the paper. Any errors and omissions remain ours.

1 Introduction

Outliers – observations that "appear to deviate markedly from other members of the sample in which they occur" (Grubbs, 1969) – are omnipresent in real-world data sets, and are almost always a cause for concern. If the 'abnormality' of extreme values is the result of measurement error, rather than a genuine manifestation of the variability of the data, then including outliers in calculations can cause serious bias to many statistics of interest.

Given the stakes, it is no surprise that the task of dealing with outliers often takes up considerable time and effort on the part of the analyst (Mancini and Vecchi, 2022). Even the very foundation of any strategy to handle outliers – identifying which observations qualify as 'extreme' – is far from trivial. Underlying a number of popular outlier detection methods is a basic algorithm: the analyst defines some concept of distance from the bulk of the distribution, which then allows to flag as 'extreme' any observation for which this distance is larger than a certain threshold. This principle is as simple as it is elusive: in practice, it is declined in a great many different ways, and the criteria used are not always well defined. This poses a number of problems: 'good practices' fail to consolidate, and the steps taken to deal with outliers are often difficult to replicate, which is an issue for scientific rigor as well as for comparability of data analysis.

In this paper we present `outdetect`, a tool designed to help the analyst *i)* identify extreme values in a univariate distribution based on a transparent procedure, and *ii)* gauge their potential impact on selected statistics of interest.[2] While the focus is on distributional analysis (*e.g.* inequality and poverty estimation – see Cowell and Flachaire, 2007, 2015), the use of `outdetect` naturally extends to any situation where the presence of 'too large' or 'too small' values in a distribution is an issue.

The algorithm employed by `outdetect` relies on the normalization of the target variable, and the imposition of cutoffs to define an outlier region. A popular version of this procedure is to log-transform the distribution, and flag observations that are more than two or three standard deviations from the mean (*e.g.*, Deaton and Tarozzi, 2005). However, `outdetect` allows users to choose from a number of alternative (and more flexible) normalizing transformations, and to use a number of alternative and statistically robust measures of location and scale. The use of transformations other than the log allows the analyst to apply a consistent detection criterion to a wide range of variables, which is

[2] The command `outdetect` can be installed in Stata by typing `ssc install outdetect`.

especially useful when dealing with the skewed and heavy-tailed distributions that are commonplace in welfare analysis (*e.g.* household consumption, income, and wealth). Some of these techniques are not yet available as stand-alone Stata commands (namely the transformation proposed in Yeo and Johnson, 2000). Robust location and scale estimators are useful when the detection rule itself is sensitive to precisely those extreme values it is designed to flag (Davies and Gather, 1993).[3]

`outdetect` does *not* offer an automated way to treat outliers – to replace or drop them – and deliberately so. Any alteration of the raw data is potentially problematic, and must be informed by careful investigation of the nature of extreme values. Instead, `outdetect` focuses on sensitivity: it produces an array of statistics using both the 'raw' distribution (the data as they are) and its outlier-free counterpart (a distribution where observations that are flagged as outliers have been excluded from all calculations). Borrowing from the analytical framework developed by Hampel (1974) and Hampel, Ronchetti, Rousseeuw and Stahel (1986), it also produces two types of diagnostic plot, the Incremental Trimming Curve (ITC) which plots the value of a statistic of interest against the proportion of extreme values that are trimmed from the sample, and the High-influence Observation Curve (HOC) first suggested by Cowell and Flachaire (2007), which describes the effect of any one (extreme) observation on the estimated value of the statistics. These instruments allow the analyst to assess the influence of extreme values on results, and inform next steps.

Of particular note is the fact that `outdetect` allows for the use of complex survey settings, so that the aforementioned comparisons can be performed using population statistics.

In the last part of the paper we use `outdetect` to investigate the influence of extreme values on key inequality and poverty statistics, computed on the basis of household budget survey data. We use a collection of data sets from the Rural Livelihoods Information System (RuLIS), a joint initiative of the Food and Agriculture Organization (FAO), the World Bank, and the International Fund for Agricultural Development (IFAD). We find that the share of observations flagged as outliers of per capita consumption is 0.8% on average, never exceeding 2.5%; that the presence or inclusion or exclusion of these observations from calculations causes differences of as many as 6 percentage points for the

[3] Commands `bacon` (Billor, Hadi and Velleman, 2000; Weber, 2010), and `gboxplot` (Bruffaerts, Verardi and Vermandele, 2014; Verardi and Vermandele, 2018) are also available to detect outliers with Stata, though their functionalities and range of application are adjacent, rather than overlapping, with `outdetect`.

Gini index, circa 10 for the Mean Log Deviation and the Atkinson index, and 29 for the Theil index. Poverty indices are found to be less sensitive. While these findings are, of course, specific to the surveys that were included in the exercise, they give empirical support to more general considerations, with important implications for the practice of applied welfare measurement, in particular for welfare comparisons: a common framework for detecting and treating extreme values is essential for both international comparisons and within-country time trends.

The paper is organized as follows: section 2 illustrates the method for outlier detection that is used in the rest of the paper; section 3 presents the command `outdetect`; section 4 shows international evidence on the sensitivity of poverty and inequality estimates to extreme values of welfare indicators; section 5 offers some concluding remarks.

2 Outlier detection

In this section the variable of interest – the target for outlier detection – will be denoted by x, and its probability density function by $f(x)$. For expositional simplicity, we shall refer to x as 'consumption', but nothing prevents one from thinking of x as standing for price, unit value, quantity, wage, income, any expenditure component, or any other continuous variable whose extreme values are seen as potentially problematic.

How exactly should one identify outliers in the distribution of x? When does a high or low value of consumption qualify as 'extreme', that is, 'too far away' from the bulk of the distribution, so much as to arise suspicion as to whether it is genuine?

Out of the many criteria proposed in the literature, `outdetect` uses one that is based on the construction of an *outlier region* with reference to $f(x)$ (Davies and Gather, 1993; Gather and Becker, 1997). If the distribution of interest is known – for example, if $f(x)$ is Normal – then one can consider an observation to be an outlier if it falls into a range of values that occur with arbitrarily low probability. An observation x_i $(i = 1, \dots, n)$ falling into a range defined in this way could conceivably be produced by the theoretical distribution $f(x)$, but that would be a rare occurrence, making x_i an extreme value, or an outlier. A conventional application of this criterion identifies the bounds of the outlier region for a Normal distribution as the mean, μ, plus or minus three times the standard deviation, σ (each tail region defined in this way has a probability of about 1%).

There are two issues with the application of this procedure to most situations. First, the distribution: $f(x)$ is not known, and certainly the empirical distribution of consumption, and most other variable of interest to welfare analysts, is not Normal. Rather, it is typically *asymmetric* and *heavy-tailed* compared to the Gaussian distribution. However, if one can transform x into something that is approximately Normal, the algorithm can still be applied: observations that are flagged in the transformed distribution are also outliers of the untransformed distribution. A second, subtler problem is related to the definition of the outlier detection region in terms of mean and standard deviation of the distribution. The empirical mean and standard deviation are not robust statistics, *i.e.* they are vulnerable precisely to the outliers one is concerned about.

These considerations suggest a more general outlier detection strategy, which is implemented by `outdetect`. It can be broken down in two steps:

1) transform the variable of interest to induce *normality* in its empirical pdf;
2) set *robust* thresholds to identify the outlier region.

To accomplish the first step, one needs to select an appropriate normalizing transformation $g(\cdot)$. We shall denote the transformed (normalized) variable as $y = g(x)$. Section 2.1 elaborates on the transformations available in `outdetect`. To accomplish the second step, one needs to pick a measure of central tendency, or location (such as the mean, or a robust alternative), and a measure of dispersion, or scale (such as the standard deviation, or a robust alternative). Section 2.2 elaborates on the measures of location and scale that are available in `outdetect`.

The rule used to detect outliers can be expressed conveniently in terms of the *z-score* of y, defined as $z = (y - \mu)/\sigma$. Here, the letters μ and σ indicate the mean and standard deviation of y, or any robust alternatives (for convenience, we shall continue to refer to z as a *z-score*, albeit a 'robust' one, when we depart from the mean and standard deviation in favor of robust measures). The goal is to choose a conventional value z_α to define an outlier region over the distribution of z, as follows:

$$z_i = \left|\frac{y_i - \mu}{\sigma}\right| > z_\alpha \qquad (1)$$

According to equation (1), an observation of the variable of interest, x_i, is flagged as an outlier if z_i – the (robust) *z-score* associated with the transformation $y_i = g(x_i)$ – exceeds,

in absolute value, the z_α quantile of the distribution of *z-scores*.[4] Equivalently, the value x_i is flagged as an outlier if its transformation, y_i, falls outside the following region:

$$[\mu - z_\alpha \times \sigma\,, \mu + z_\alpha \times \sigma] \tag{2}$$

As for the choice of z_α, the threshold for the outlier region in terms of the (robust) *z-score*, a degree of arbitrariness is inevitable: z_α is not tied to any statistical requirement, beyond the need to identify a 'low-probability' outlier region. The choice of the value 3 is customary, but smaller and larger values (2.5, 3.5, or 4) are not uncommon. Higher values of z_α will shrink the outlier region, lower the probability thresholds for the tails of the distribution of *z-scores*, and therefore flag fewer outliers.

Once outliers are detected according to the rule described by equations (1) and (2), `outdetect` produces a table that shows the numbers of observations that were flagged, both at the 'top' and at the 'bottom' (*i.e.* large and small extreme values, respectively), and their proportion over the total number of observations. The output also includes a set of summary statistics and diagnostic tools, designed to inform the user about the sensitivity of the statistics of interest to the presence of outliers. These are described in Section 2.3.

2.1 Normalizing transformations

There is no lack of choice of transformations for achieving approximate normality in the literature. During the early 2000s, in a contribution to the 'great Indian poverty debate', Deaton and Tarozzi (2005) have explored the use of the natural logarithm as a normalizing transformation for unit values of commodities consumed by households. This is the simplest of transformations – the logarithm "squeezes" large values more than small ones, so that the logarithm of a skewed distribution becomes more symmetrical, and closer to a Normal. Dupriez (2007) followed another route, and adopted the Box-Cox transformation (Box and Cox, 1964), which includes the log transformation as a special case. Other useful transformations are available, such as those proposed in Yeo and Johnson (2000), and Friedline et al. (2014), among others. Table 1 lists the transformations that are implemented

[4] The use of the absolute value in equation (1) allows for the detection of both 'top' and 'bottom' outliers (both 'too large' and 'too small'), and can be easily replaced by one-sided versions when the focus is only on one tail of the distribution.

by `outdetect`, with x indicating the target variable (say, consumption) and y indicating the transformed variable (say the log-consumption).

Table 1. Transformations to approximate normality

Description	Formula for y	Source
Natural log	$y = \ln(x)$	1
Log base 10	$y = \log_{10}(x + a)$ with $a = \max[0, -(\min(x) - 0.001)]$	2
Box-Cox	$y = \begin{cases} \frac{x^\lambda - 1}{\lambda} & \text{if } \lambda \neq 0 \\ \ln(x) & \text{if } \lambda = 0 \end{cases}$	3
Yeo-Johnson	$y = \begin{cases} \frac{x^\lambda - 1}{\lambda} & \text{if } x \geq 0, \lambda \neq 0 \\ \log(x+1) & \text{if } x \geq 0, \lambda = 0 \\ \frac{-\left[(-x+1)^{2-\lambda} - 1\right]}{2-\lambda} & \text{if } x < 0, \lambda \neq 2 \\ -\log(-x+1) & \text{if } x < 0, \lambda = 2 \end{cases}$	4
Inverse hyperbolic sine	$y = \ln(x + \sqrt{x^2 + 1})$	5
Square root	$y = \sqrt{x}$	6

Sources: [1,2 and 6] any math textbook; [3] Box and Cox (1964); [4] Yeo and Johnson (2000); [5] Friedline et al. (2014).

Each of the transformations in Table 1 has properties that fit different needs. For example, while the natural log may only be applied to strictly positive variables, $\log_{10}(x + a)$ takes care of negative and zero values, too. Similarly, while the use of Box-Cox is limited to strictly positive variables, the Yeo-Johnson and the inverse hyperbolic transformations apply to all variables and perform relatively better in the presence of highly skewed distributions (which makes them particularly suitable for the analysis household wealth data). A general criterion to select the "best" transformation is goodness-of-fit: the transformation that provides the best approximation to a Normal distribution for the specific variable and data set in use will be the best choice for that particular context. `outdetect` allows the user to maximize goodness-of-fit relying on the Pearson chi-squared test (Snedecor and Cochran, 1989). As pointed out by Peterson and Cavanaugh (2019), the Pearson statistic, P, divided by its degrees of freedom, df, converges to 1 when the data approaches a Gaussian distribution: P/df can be interpreted as a measure of how close a distribution is to normality, and used to rank transformations according to how successful they are in normalizing the data.

2.2 Measures of location and scale

When μ is the mean and σ is the standard deviation of the normalized consumption distribution y, the outlier detection rule in equations (1) and (2) *itself* is sensitive to the presence of outliers. Davies and Gather (1993) suggest the use of robust location and scale measures to contrast this problem. There is little dispute on the use of the *median* instead of the mean: the median is simple to calculate, and, unlike the mean, provides a greater degree of resilience to the presence of outliers.[5] On the other hand, the choice of a robust scale estimator appears to be somewhat debated. Table 2 shows a selection of candidate estimators for the σ parameter in equations (1) and (2).

Table 2. Robust scale estimators for the transformed variable *t*

Estimator	σ	Source
IQR	$IQR = Q3 - Q1$	1
MAD	$MAD = 1.4826 \times med\lvert y_i - med[y]\rvert$	2
S	$S = 1.1926 \times med_i\{med_j\lvert y_i - y_j\rvert\}$	3
Q	$Q = 2.2219 \times \left(\lvert y_i - y_j\rvert; i < j\right)_{(k)}$	4

Note: $Q3$ and $Q1$ denote the 75th and 25th percentiles, respectively; t is the transformation of the target variable x. Sources: [1] any statistics textbook, [2] Hampel (1974); [3, 4] Rousseeuw and Croux (1993). The use of (k) as a subscript indicates that the data have been sorted in increasing order: given a sample of n observation, the k-th order statistic of the sample is its k-th smallest value.

An outlier detection algorithm based on a *z-score* with the interquartile range (IQR) as the denominator has been experimented with in Dupriez (2007), for instance. Hampel (1974) suggested the use of the *median absolute deviation* (MAD), defined in row 2 of Table 2: the MAD is the median value of the distance between the transformed expenditure of each household (y_i) and the center of the distribution, as estimated by the median of the transformed distribution ($med[y]$). Under the assumption that y is close enough to the

[5] The price for the robustness of the median is a loss of efficiency with respect to the mean (Rousseeuw and Hubert 2017: 2), but in large samples the main, if not the only, property that matters is their bias – efficiency is only relevant for small samples. Note that we are assuming unweighted estimators. The *weighted* median does not, in fact, qualify as a robust statistic (Filzmoser, Gussenbauer and Templ 2016: 15).

standard Normal distribution, then the coefficient 1.4826 in the formula is required for making the MAD an unbiased estimator of the standard deviation. Rousseeuw and Croux (1993) introduced the *S* and *Q* estimators, described in rows 3 and 4 of Table 2, which are shown to have the same 50% maximal breakdown point of MAD, but be more efficient than MAD when $f(x) \sim N(\mu, \sigma)$. Moreover, *S* and *Q* are location-free scale estimators which makes them suitable to deal with skewed distributions, the case of highest interest to welfare analysts, thanks to their location-free nature.

Both *S* and *Q* are "sturdy" estimators of the standard deviation, but both estimators are computationally demanding.[6] Nevertheless, efficient algorithms proposed by Croux and Rousseeuw (1992) make their burden manageable. Given that its core is written in Mata, `outdetect` takes full advantage of these improvements, speeding up the computation considerably.[7]

The availability of alternative estimators for the scale parameter of y begs the same question that comes up for normalizing transformations: which one should the analyst pick? No alternative is clearly superior according to the literature. Rousseeuw and Croux (1993) show that *S* and *Q* have desirable statistical properties, and once the burden associated with their computation is reduced – `outdetect` is fast, due its reliance on Mata – they turn out to work well in most practical applications. For practical purposes, the analyst may refer to the following ranking: $Q \succcurlyeq S \succ MAD \succ IQR$, where we use the sign "$\succ$" to mean 'is preferred to' and "$\succcurlyeq$ for 'weakly preferred to'.[8]

[6] To calculate the S estimator, for example, one needs to compute, for each household i in the sample, the expression $\{med_j|y_i - y_j|\}$ for j=1, …, n. This gives n numbers, the median of which gives the estimate S (the number 1.1926 in the formula is required for making S a consistent estimator of the standard deviation under the assumption of normality). Similarly, the Q estimator (row 4 in Table 2) is obtained by sorting all pairwise distances $|y_i - y_j|$ and taking the value that occupies the k-th position in the ranking, with k being roughly half the number of observations (the number 2.2219 in the formula is required for making Q a consistent estimator of the standard deviation under the assumption of normality).

[7] In order to compute the *Q* statistic, `outdetect` adapts the code written by Ben Jann for the `robstat` command (Jann, Verardi and Vermandele, 2018).

[8] The final choice between S and Q is subjective, because their advantages and disadvantages are not easily compared (Rousseeuw and Croux, 1992). Indeed, while they share the nature of location-free robust scale estimator with a 50% breakdown point, Q is slightly preferred in terms of efficiency at the Gaussian model and, unlike S, it has a continuous influence function. On the other hand, S requires only half as much computation time and storage as Q.

2.3 Sensitivity of statistics to the presence of outliers

Once outliers have been identified, `outdetect` assesses the sensitivity of selected statistics to whether or not those same outliers are included in calculations. This is done by taking two heuristic approaches (Hampel 1974). On the one hand, the output of `outdetect` reports comparisons between estimates based on 'raw' data (that is, inclusive of all values) and data where all observations classified as outliers are excluded from calculations. On the other hand, `outdetect` includes an option to draw the Incremental Trimming Curve (ITC), which shows how the value of a statistic of choice changes as the largest or smallest observations in the data set are consecutively excluded from calculations. The ITC is inspired by the (finite-sample versions of) the empirical influence curve (Hampel 1974, Hampel, Ronchetti, Rousseeuw and Stahel 1986, ch. 2), and Tukey's sensitivity curve (Huber 2002), and similar to the curve proposed by Cowell and Flachaire (2007) to identify influential observations.

A first set of statistics that appear in the output are standard descriptive statistics: the mean and the median, the standard deviation (abbreviated with SD), the coefficient of variation (CV), and the interquartile range (IQR). A second set of statistics focuses on inequality measures and Foster, Greer, and Thorbecke (1984) poverty measures.

As for the ITC, it is defined as follows:

$$\widehat{ITC}\left(x_{(i)}\right) = \hat{J}_{(i)} \;\; for\; i = 0, 1, \dots, I$$

where $\hat{J}_{(i)}$ denotes the statistics of interest calculated on the distribution of x, after sorting the values of x and excluding the i-th *cumulated* observation. If $\hat{J}$ denotes, say, the Gini index, then $\hat{J}_{(i)}$ is the Gini index for x, calculated leaving out of the sample the first i observations. If $i=0$, then $\hat{J}_{(0)} = \hat{J}$, which corresponds to the case when no extreme value is discarded, and the ITC returns the Gini index estimated on the raw data set. If $i=1$, then $\hat{J}_{(1)}$ is the value of the Gini index obtained after discarding the first extreme value. Similarly, $\hat{J}_{(2)}$ corresponds to the Gini index calculated when the two most extreme values have been discarded from the data set. Note that data can be thought of as sorted either ascending or descending, so that `outdetect` produces two ITCs, one where the impact of 'too small' values is assessed, the other focused on the impact of 'too large' values. Overall, the gradient of the ITC curves provides a neat indication of the extent to which the chosen statistics is affected by the presence of extreme values: the steeper the ITC, the higher the impact. The ITC is further discussed and illustrated in sections 3 and 4.

3 How to use the `outdetect` command

This section is devoted to illustrating the typical use of `outdetect`. After explaining the syntax, we provide examples of the command "in action", using a sample of 12,447 households from Malawi's Fourth Integrated Household Survey (2016-2017).[9]

`outdetect` can be installed from the Statistical Software Components Archive by typing `ssc install outdetect`. Stata 15.1 is the earliest version that can run `outdetect`. The core of the command is written entirely in Mata to minimize computation time. The command handles complex survey settings automatically, when `svyset` is used to declare the sampling design features of the data to Stata (see `help svyset`). The use of `pweights` is also allowed. The general syntax of the command is as follows:

`outdetect` *varname* [*if*] [*in*] [*weight*] [, *options*]

`outdetect` identifies extreme values, either "too small" or "too large" observations, in the distribution of `varname`. We shall call these observations bottom outliers and top outliers, respectively (small values being at the bottom of the distribution of `varname`, and large values being at the top). By default, `outdetect` creates a new variable, `_out`, containing numeric codes that flag outliers of `varname`:

0	observation is not an outlier
1	observation is a bottom outlier ("too small")
2	observation is a top outlier ("too large")

The output of `outdetect` reports "Raw" statistics (computed using `varname` as is), as well as "Trimmed" statistics (computed using just those observations of `varname` that are not flagged as outliers). A full description of all the available options is provided in the `outdetect` help file.

3.1 Basic usage

In this example, the target variable is per capita annual household expenditure (here called `pce`) in Malawi (monetary amounts are expressed in thousands of Malawian kwacha

[9] The data set is part of the Rural Livelihoods Information System (RuLIS), and is also publicly available at the World Bank Microdata Library.

(MWK), with 1 USD exchanged against circa 800 MWKs at the time of writing). As a first step, we load the data into memory, specify survey settings, and summarize `pce`:

```
use malawi, clear

summarize pce [aw=weight], detail
```

```
                    Per capita expenditures
-------------------------------------------------------------
      Percentiles      Smallest
 1%           170             45
 5%           253             58
10%           314             61       Obs               12,447
25%           441             70       Sum of wgt.   16307879.5

50%           649                      Mean                 898
                        Largest       Std. dev.          3,926
75%           991          15,678
90%         1,513          18,704       Variance       15414949
95%         2,080         147,453       Skewness             81
99%         3,954         340,687       Kurtosis          6,877
```

The distribution of `pce` displays the typical features of expenditure distributions from survey data anywhere: it is highly skewed to the right, heavily leptokurtic, and some values appear to be abnormally high, as well as, possibly, abnormally low, with respect to the bulk of observations. The two largest observations in the sample, in particular, are one order of magnitude above any other large values in the distribution. Similarly, the smallest values in the sample (corresponding to 20-30 cents of US dollar per day) appear implausibly low. To gain a better sense of the extent that these extreme values might impact the analysis, you can issue `outdetect` using the default syntax (note that while `outdetect` allows users to specify weights the same way as other Stata commands, `svy`-setting the data set prior to running `outdetect` is the recommended practice):

```
svyset psu [pweight = weight]

outdetect pce
```

outdetect **set-up**:

Normalization: Yeo and Johnson (2000)
Z-score: (x-median)/q
α = 3
Outlier detection target: top and bottom
(12447 observations are used)

Incidence of outliers:

	Freq.	Percent	Share
Bottom	**82**	**0.66**	**67.77**
Top	**39**	**0.31**	**32.23**
Total	**121**	**0.97**	**100.00**

Statistics for raw and trimmed pce:

	Raw	Trimmed
Summary stats		
Mean	**897.88**	**828.38**
Median	**648.73**	**648.35**
SD	**3926.19**	**641.13**
CV(%)	**437.27**	**77.40**
IQR	**550.12**	**543.91**
Inequality		
Gini	**0.4063**	**0.3565**
MLD	**0.2839**	**0.2076**
Theil	**0.5057**	**0.2232**
CV2	**9.5596**	**0.2995**
A(0.125)	**0.0541**	**0.0273**
A(1)	**0.2472**	**0.1875**
A(2)	**0.3776**	**0.3207**
p90/p10	**4.8185**	**4.7558**

The outcome of `outdetect` is organized into three parts. The top panel specifies the settings of the outlier detection procedure, that is, the ways in which the parameters of equation (1) (y, μ, σ, and z_α), are set, and whether outliers are detected in both tails of the distribution. When the default syntax is used, `outdetect` uses the Yeo and Johnson (2000) transformation. In this example, μ is the median of y, σ is the Q estimator, and z_α is equal to 3. Because the *z-score* in equation (1) is considered in absolute value, outliers are detected both at the top and bottom of the distribution (*i.e.* both large and small values).

The middle section of the output summarizes how many observations are flagged as outliers. The first column (Freq.) reports the frequency: 121 observations in total, 82 at the

top and 39 at the bottom; the second column (Percent) shows that observations flagged as outliers amount to 0.97 percent of the total sample size, with a prevalence of bottom (0.66 percent) over top (0.31 percent) outliers; the third column (Share) gives the breakdown by bottom and top outliers. In our example, about 2 out of 3 outliers are 'bottom outliers'. Overall, the information on the distribution of the outliers between the two tails of the distribution helps analysts to form expectations on the potential instability of their statistics of interest due to the presence of extreme values. Cowell and Flachaire (2007), for instance, provide a full account of the behavior of inequality measures in the presence of extreme values: according to their Result 1, Generalized Entropy indices with coefficient $\theta > 1$ (*e.g.* the squared coefficient of variation) are very sensitive to high incomes in the data, while Result 2 shows that Atkinson measures with $\varepsilon > 1$ (where ε denotes the inequality aversion parameter) are very sensitive to low incomes in the data.

The default categorical `_out` variable flags the two types of outliers in the data set: `_out` takes on the value of 0 if the observation is not an outlier, 1 if it is a bottom outlier, and 2 if it is a top outlier. This can be verified by issuing `tabulate _out`.

The bottom section of the output contains the core results of `outdetect`: it compares an array of 16 descriptive statistics obtained "with" and "without" outliers. The first column ("Raw") uses `pce` as is, meaning that the statistics are computed using all nonmissing observations of `pce`; the second column ("Trimmed") excludes all observations flagged as outliers from calculations. In the case of Malawi, although the incidence of observations flagged as outliers is small (0.97%), their impact on most indicators is quite significant.[10]

In certain cases, as for the variable considered here, it can be useful to gauge the sensitivity of poverty estimates to extreme values, as well. The user can expand the output by specifying a poverty line, as shown in the following image (the `nogen` option is used so the command refrains from creating the default `_out` variable, which currently already exists in the data set, given that `outdetect` has been issued before). The additional statistics at the bottom of the output indicate that poverty estimates are not as impacted by the exclusion of extreme values.

[10] The addition of standard errors and tests to assess the statistical significance of differences is a feature of an upcoming update of `outdetect`.

```
outdetect pce, pline(300) nogen
```

outdetect **set-up**:

Normalization: Yeo and Johnson (2000)
Z-score: (x-median)/q
α = 3
Outlier detection target: top and bottom
(12447 observations are used)

Incidence of outliers:

	Freq.	Percent	Share
Bottom	**82**	**0.66**	**67.77**
Top	**39**	**0.31**	**32.23**
Total	**121**	**0.97**	**100.00**

Statistics for raw and trimmed pce:

	Raw	Trimmed
Summary stats		
Mean	**897.88**	**828.38**
Median	**648.73**	**648.35**
SD	**3926.19**	**641.13**
CV(%)	**437.27**	**77.40**
IQR	**550.12**	**543.91**
Inequality		
Gini	**0.4063**	**0.3565**
MLD	**0.2839**	**0.2076**
Theil	**0.5057**	**0.2232**
CV2	**9.5596**	**0.2995**
A(0.125)	**0.0541**	**0.0273**
A(1)	**0.2472**	**0.1875**
A(2)	**0.3776**	**0.3207**
p90/p10	**4.8185**	**4.7558**
Poverty		
H	**0.0841**	**0.0823**
PG	**0.0186**	**0.0172**
PG2	**0.0062**	**0.0053**

Poverty line: 300

The results displayed so far are, of course, dependent on the settings of the outlier detection routine, which users may wish to customize. For instance, to reproduce the procedure originally implemented by Deaton and Tarozzi (2005), they will specify that *i)* the normalization of `pce` be the natural logarithm, *ii)* the *z*-score be computed by subtracting the mean and dividing by the standard deviation of the log of `pce`, *iii)* outliers be detected only at the top of the distribution, and *iv)* the threshold marking the outlier region set at 2.5

instead of 3 (the default value used by `outdetect`). The following code implements these settings:

```
outdetect pce, norm(ln) zscore(mean std) out(top) alpha(2.5) replace
```

outdetect **set-up**:

Normalization: natural logarithm
Z-score: (x-mean)/std
α = 2.5
Outlier detection target: top
(12447 observations are used)

Incidence of outliers:

	Freq.	Percent	Share
Bottom	**0**	**0.00**	**0.00**
Top	**157**	**1.26**	**100.00**
Total	**157**	**1.26**	**100.00**

Statistics for raw and trimmed pce:

	Raw	Trimmed
Summary stats		
Mean	**897.88**	**786.32**
Median	**648.73**	**641.44**
SD	**3926.19**	**521.38**
CV(%)	**437.27**	**66.31**
IQR	**550.12**	**533.24**
Inequality		
Gini	**0.4063**	**0.3328**
MLD	**0.2839**	**0.1814**
Theil	**0.5057**	**0.1840**
CV2	**9.5596**	**0.2198**
A(0.125)	**0.0541**	**0.0227**
A(1)	**0.2472**	**0.1659**
A(2)	**0.3776**	**0.2985**
p90/p10	**4.8185**	**4.6065**

The `replace` option generates a new `_out` variable, which replaces the existing one.

To apply the best fitting transformation, the user may specify option `bestnormalize`:

```
outdetect pce, bestnormalize replace
```

outdetect **set-up**:

Normalization: Box and Cox (1964)
Z-score: (x-median)/q
α = 3
Outlier detection target: top and bottom
(12447 observations are used)

Incidence of outliers:

	Freq.	Percent	Share
Bottom	**82**	**0.66**	**67.77**
Top	**39**	**0.31**	**32.23**
Total	**121**	**0.97**	**100.00**

Statistics for raw and trimmed pce:

	Raw	Trimmed
Summary stats		
Mean	**897.88**	**828.38**
Median	**648.73**	**648.35**
SD	**3926.19**	**641.13**
CV(%)	**437.27**	**77.40**
IQR	**550.12**	**543.91**
Inequality		
Gini	**0.4063**	**0.3565**
MLD	**0.2839**	**0.2076**
Theil	**0.5057**	**0.2232**
CV2	**9.5596**	**0.2995**
A(0.125)	**0.0541**	**0.0273**
A(1)	**0.2472**	**0.1875**
A(2)	**0.3776**	**0.3207**
p90/p10	**4.8185**	**4.7558**

In this case, the "best" normalization turns out to be Box and Cox (1964); however, results in terms of outliers flagged do not change with respect to those obtained using the default Yeo and Johnson (2000) transformation since, the latter, is equivalent to the generalized Box and Cox (1964) $[(x^{\lambda}+1)-1]/\lambda$, for $x > -1$, where the shift constant 1 is included.

Finally, the user can specify option `excel`, to export the table of results in an Excel file which is saved in the current working directory or in any other specified location:

```
outdetect pce, excel(demo, replace) replace
```

In this case, the `replace` option within parentheses refers to the Excel file, while the one outside refers once again to the default `_out` variable.

3.2 Diagnostic graphs

The sensitivity of the statistics of interest to the presence of extreme values can also be assessed independently of the settings of the outlier detection procedure, by producing one or more diagnostic graphs. Depending on the graph selected, the output shown on the screen may change.

First is the Incremental Trimming Curves (ITC), defined in section 2.3. For example, the ITC for a selection of statistics of interest can be plotted by issuing the commands below:

```
outdetect pce, graph(itc(2: mean))          (figure 1, panel a)
outdetect pce, graph(itc(2: gini))          (panel b)
outdetect pce, graph(itc(2: h pline(300))   (panel c)
outdetect pce, graph(itc(2: pg pline(300))  (panel d)
```

The syntax specifies that *(i)* the diagnostic graph to be produced is the ITC, *(ii)* the curve should be plotted for the top and bottom 2 percent of observations, and *(iii)* the statistics of interest are, the mean, the Gini index, the poverty headcount ratio, and the poverty gap index.

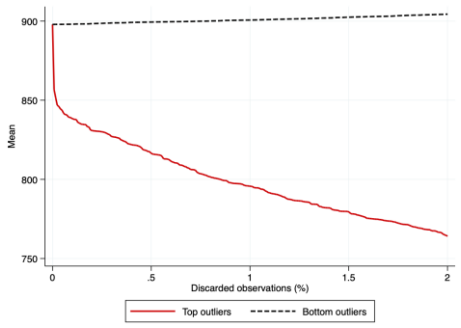
The results are shown in Figure 1. To clarify the interpretation of the curves, let us take the top-right panel, showing the ITC for the Gini index. The solid line shows the value of the Gini index as the *largest* observations are dropped from the sample, one at a time; the dashed line shows Gini when the *smallest* observations are removed, one at a time. The steeper the curve, the more sensitive Gini is to the presence of extreme values.[11]

As expected, mean expenditure turns out to be quite sensitive to the presence of large outliers (much less so for the outliers in the left tail). Similarly, the Gini index behaves consistently with the analysis of Cowell and Flachaire (2007): it shows a remarkable stability to bottom outliers, and high sensitivity to top outliers. Finally, both the headcount and poverty gap indices in panels c and d show an asymmetric response to extreme values: they are robust (but not totally insensitive) to extremely large values, but sensitive to bottom outliers. Even more so is the poverty gap squared index (not shown in the figure), consistently with its analytical properties (Cowell and Victoria Feser 1996b).

[11] Each time an observation is dropped, the weights of the remaining observations are recalibrated so as to sum up to the entire population.

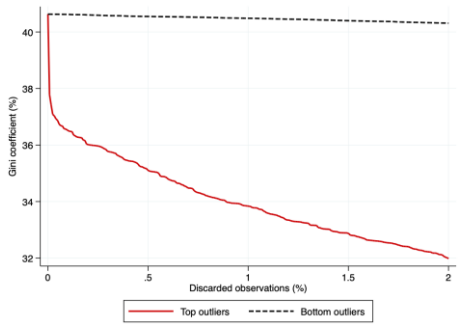
Figure 1. Incremental Trimming Curves for Malawi, 2017

a. Mean



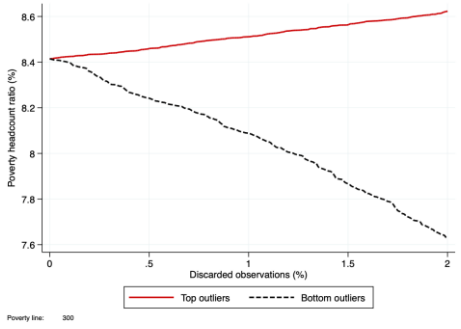
Discarded obs (%)	Mean Bottom	Mean Top
0	897.88	897.88
1	900.64	795.75
2	904.37	764.14

b. Gini index (%)



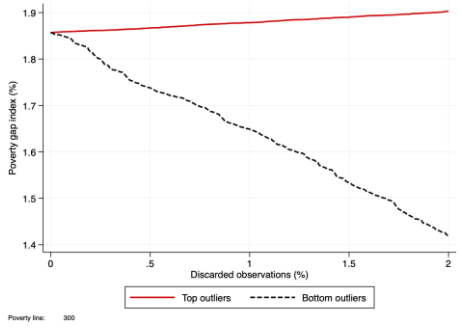
Discarded obs (%)	Gini Bottom	Gini Top
0	40.63	40.63
1	40.49	33.84
2	40.31	31.98

c. Headcount Poverty (%)



Discarded obs (%)	H Bottom	H Top
0	8.41	8.41
1	8.09	8.51
2	7.63	8.62

d. Poverty Gap Index



Discarded obs (%)	PG Bottom	PG Top
0	1.86	1.86
1	1.65	1.88
2	1.42	1.90

Note that when ITC graphs are produced, `outdetect` does not generate the default `_out` variable, nor does it show the output described in section 4.1. Instead, the command generates a table like those shown in figure 1, to facilitate the interpretation of the curve. The table reports values of the ITC for selected shares of discarded observations. For instance, if we focus on panel (a), we can interpret the table associated with the ITC for the mean as follows: when 0% of observations are discarded, the mean of `pce` is equal to

897.88 thousand MWK (per person/year); when 1% of the smallest observations are discarded, the mean is equal to 900.64 thousand MWK, whereas if 1% of the largest observations are discarded, the mean is equal to 795.75 thousand MWK; and so on.

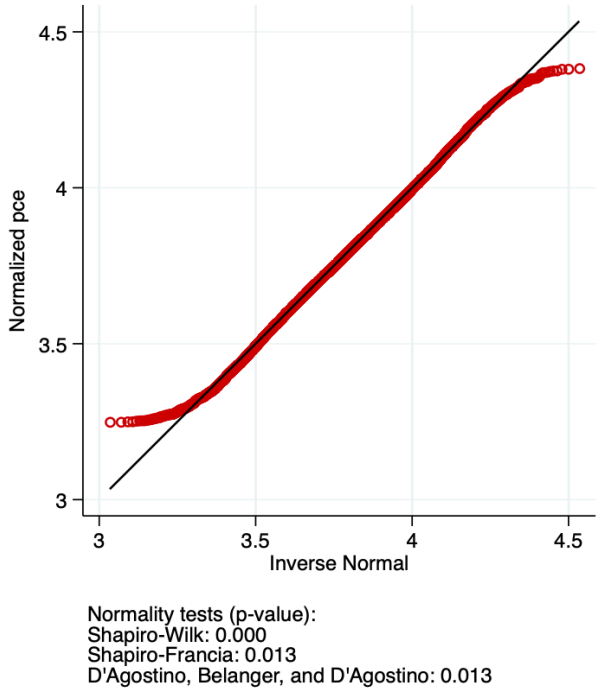
Note that users can produce the same graph by specifying that a certain *number* (not share) of observations be discarded; they will simply need to add the option `abs`, for absolute. The example below produces the ITC for the mean, focusing on the smallest and largest 10 observations in the sample:

```
outdetect pce, graph(itc(10: mean abs))
```

One last type of diagnostic plot that is available as part of `outdetect` has to do with monitoring the goodness-of-fit of the normalization of the target distribution. This is an important check because, if the normalization is successful, then the characterization of outliers as "low-probability observations", as defined by the tails of a Normal distribution, will be applied accurately to the original (non-Normal) distribution. The syntax below produces a quantile-quantile (Q-Q) plot comparing the quantiles of the transformed distribution of `pce`, which we indicated by *y* in Section 2, to those of the inverse Normal distribution having the same mean and standard deviation as *y*:

```
outdetect pce, graph(qqplot) nogen
```

Figure 2. Assessing normalization: the quantile-quantile plot



At the bottom of the QQ-plot, `outdetect` also shows the p-values from three popular normality tests: *i)* the Shapiro-Wilk test (Shapiro and Wilk 1965), *ii)* the Shapiro-Francia test (Shapiro and Francia 1972), and iii) the D'Agostino, Belanger and D'Agostino (1990) test. The null hypothesis being tested is that the variable is normally distributed, which is only rejected by the first of the three tests.

4 The influence of outliers on inequality and poverty measures

In this section, we apply the outlier detection procedure that has been illustrated in section 2 to per capita expenditure in a wide array of countries, taking advantage of the collection of survey data made available by the Rural Livelihoods Information System (RuLIS). We use data from 34 of these countries (we excluded four surveys, which were conducted in 2005 or earlier). Our sources are listed in the Appendix.

Figure 3 shows some examples of the distribution of per capita consumption aggregates based on the "raw" data.[12] All curves display a few features that are common for many variables of interest in welfare analysis (consumption expenditure, income, wealth, but also quantity of items consumed, calorie intake, unit values, and many others): they are skewed and leptokurtic (fat tails), which may be symptoms of the presence of outliers.

[12] The "raw" label is used here for convenience: survey microdata shared by National Statistical Offices have almost certainly been edited between the end of fieldwork and the time of dissemination, as is routine – but they are raw, as far as the analyst is concerned. The conversion in per capita (or adult equivalent) terms is necessary in this context, because values that may appear 'extreme' on a per-household basis may turn out not to be, once household size is accounted for.

Figure 3. Per capita expenditure (LCU/person/year), untransformed variables

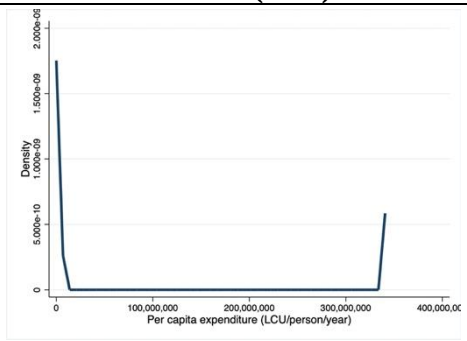
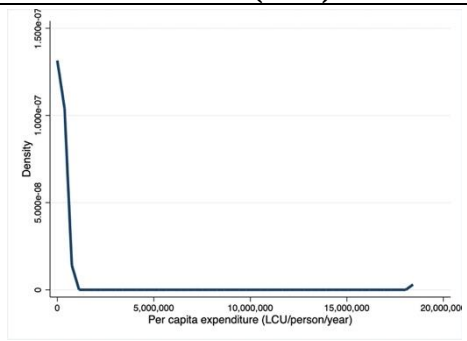
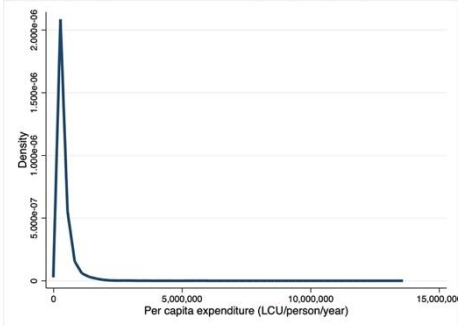
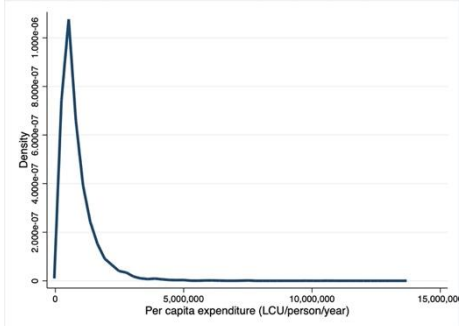
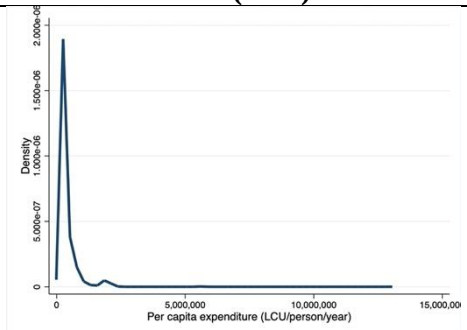
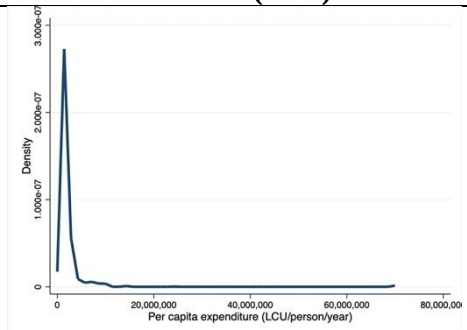
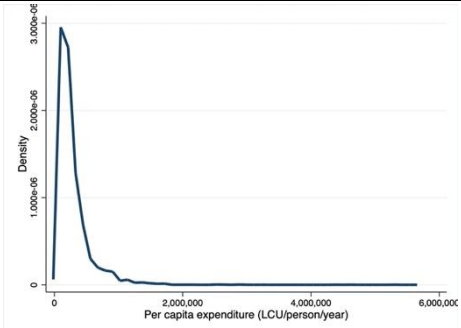
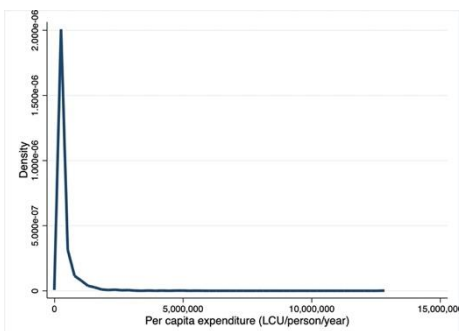
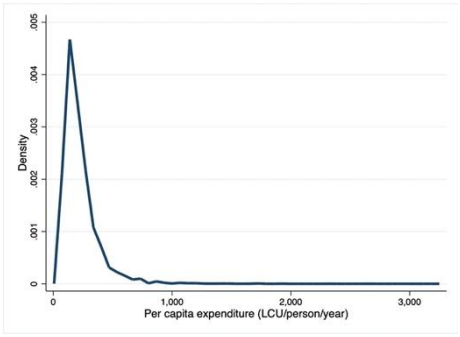
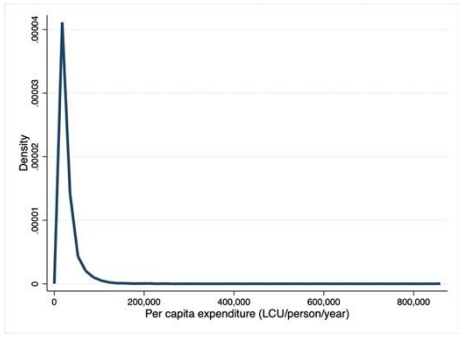
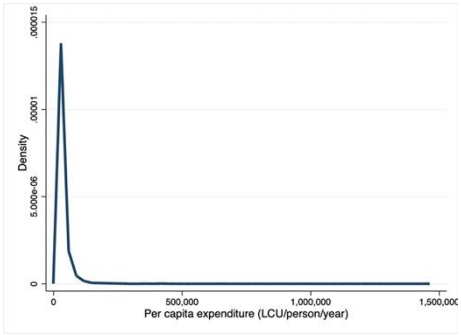
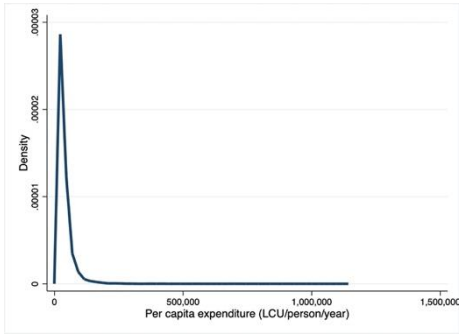
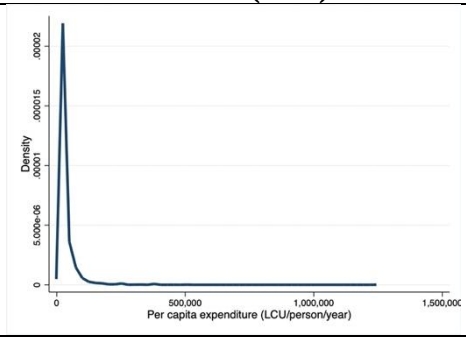
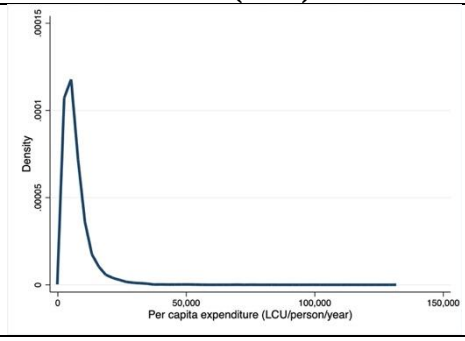
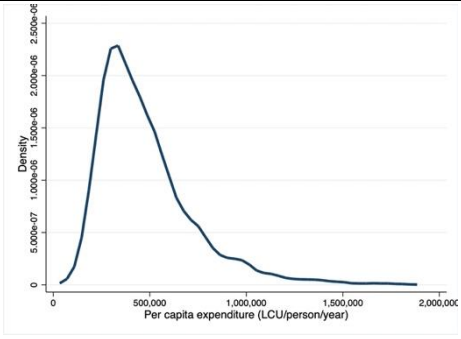
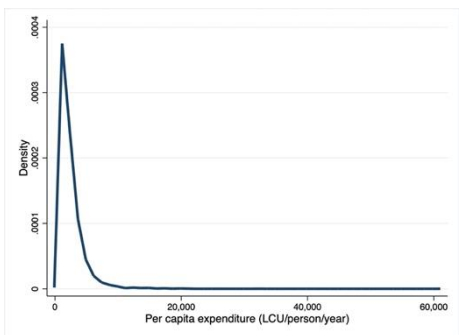
MALAWI (2017)	NIGERIA (2012)
	
CÔTE D'IVOIRE (2008)	**TANZANIA (2015)**
	
MALI (2014)	**UGANDA (2016)**
	
SENEGAL (2011)	**RWANDA (2014)**
	

Figure 3 (cont.). Per capita expenditure (LCU/person/year), untransformed variables

IRAQ (2012)	BANGLADESH (2010)
	
INDIA (2012)	**PAKISTAN (2014)**
	
MEXICO (2014)	**PERU (2015)**
	
ARMENIA (2013)	**GEORGIA (2015)**
	

Source: Authors' estimates based on data from RuLIS (2021). RuLIS database accessed on July 2021.

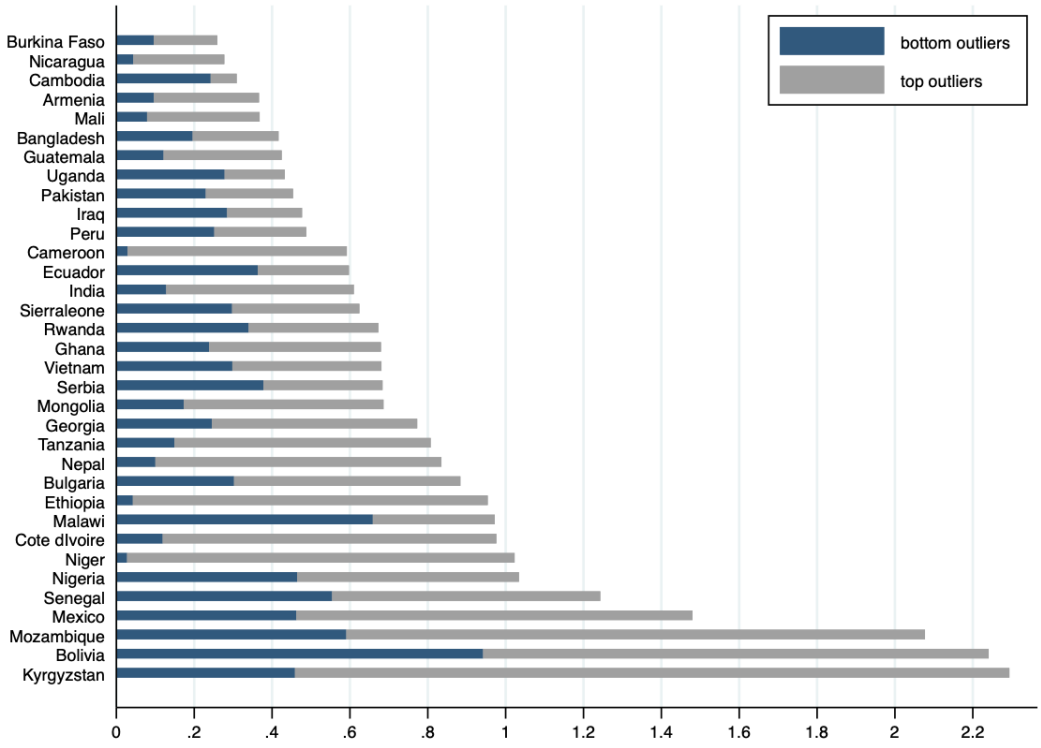
Figure 4 shows the incidence of outliers (as a percentage of the total number of observations in each sample) in the target distributions. There is remarkable heterogeneity in the fraction of flagged observations across countries; however, this fraction is relatively small: the average outlier incidence across countries is 0.8%, with a peak of 2.2% in Kyrgyzstan. The code underlying Figure 4 is the following syntax of `outdetect`:

```
svyset [pweight = weight]

outdetect pce, bestnorm
```

This implies that outliers are detected after transforming the target variable with the best fitting normalization, using the median and the *Q* estimator to compute a robust *z-score*, and applying a threshold of 3 to the resulting standardized distribution.

Figure 4. Incidence of outliers (% of sample size) of per capita consumption



Because we use a threshold of 3 to flag extreme values, the expected share of outliers detected – were the normalization of the consumption distribution perfectly successful – should be around 0.3%. The higher percentage of outliers detected is due to excess skewness of the distributions with respect to a standard normal, which is on par with the

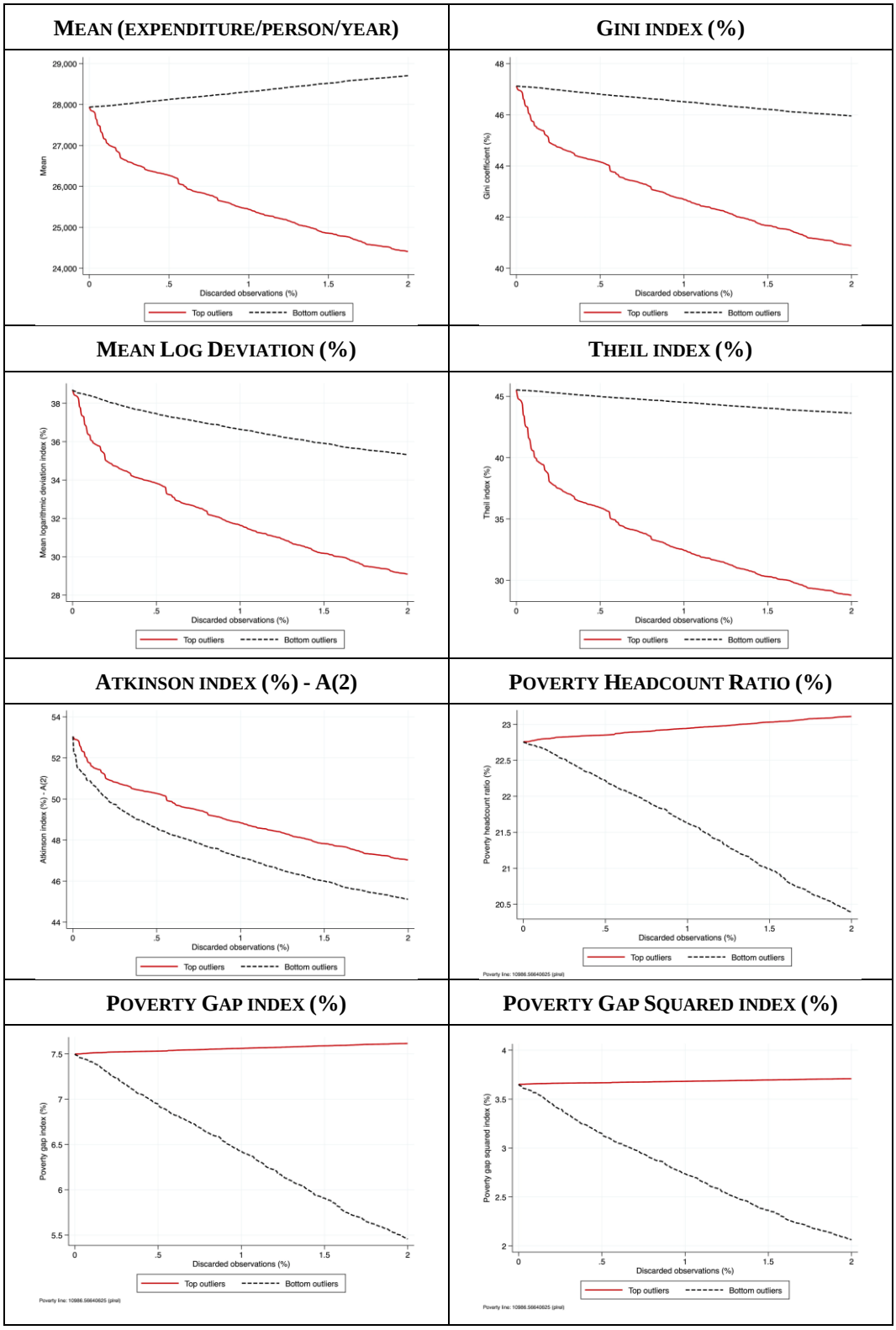
distributions displayed in Figure 3. Top outliers (extremely large observations) are more frequently flagged than bottom outliers (extremely small observations) – the latter are, however, present, despite the fact that small values typically do not attract as much attention in welfare analysis.

Besides counting how many observations are flagged, it is important to figure out how influential these observations are. A simple way to assess the sensitivity of any statistics of interest to extreme values is to track the value of the statistic – say, the Gini index – when we progressively drop the largest and smallest observations in the distribution of per capita consumption, a process we can describe as *incremental trimming* of the distribution. The resulting graph, the Incremental Trimming Curve (ITC), is displayed in Figure 5 for selected statistics.[13] The figure uses data from Mexico's Encuesta Nacional de Ingresos y Gastos de los Hogares, conducted in 2014, and reports ITCs for eight different statistics of interest: 1) average per capita consumption, 2) Gini index, 3) mean logarithmic deviation, 4) Theil index, 5) Atkinson index (with inequality aversion parameter epsilon equal to 2), 6) poverty headcount ratio, 7) poverty gap, and 8) poverty gap squared.

The ITCs in Figure 5 show that inequality statistics can be extremely sensitive to the presence of extreme values, although to different extents. The Theil index, for instance, experiences a vertical drop – 45.5% to 36% – when as few as 0.5% of large values, or about 100 observations, are excluded from calculations. The Atkinson index with a relatively high value for the inequality aversion parameter (epsilon is set equal to 2 in the figure), on the other hand, is more sensitive to the exclusion of small values. Poverty indices experience smaller and more linear changes in general when extreme values are dropped. The case of Mexico is not peculiar: in fact, it is representative of patterns that emerge in most of the countries examined in this section.

[13] The ITC is defined in Section 2.3.

Figure 5. Incremental Trimming Curves (ITCs) for selected statistics, Mexico 2014



Source: Authors' estimates based on data from RuLIS (2021).

The empirical analysis conducted so far confirms what is known from the theoretical literature: outliers are, in general, highly influential for most statistics of interest. A more operational indication of the influence of extreme values on final estimates can be derived from a simple comparison of the "raw" data ('raw' as defined in footnote 7) to some counterfactual, "what if" scenario. A *prima facie* approach is to exclude outliers from the distribution of the target variable – that is, all values flagged by `outdetect` are "rejected" (to use terminology that is common in the outlier literature), and excluded from the calculation of final estimates. Using the outlier-free distribution of consumption as a counterfactual to the raw data amounts to looking at what would happen if the observations flagged by the algorithm did not occur at all in the distribution. This is compatible with the view that observations that are flagged as outliers are *contaminants* of the true distribution (originated, for instance, from measurement error). This framework, of course, does *not* imply that trimming extreme values is the best way to treat them. In fact, the perspective in the present context is exploratory in nature: the main purpose is that of experimenting with a limit, hypothetical scenario, one where all outliers flagged are, in fact, the result of measurement error.

Table 3. Difference (in percentage points) between statistics computed before and after the exclusion of outliers

Country	Gini	MLD	Theil	Atkinson	H	PG	PG2
Armenia	0.3	0.4	0.3	0.8	0.2	0.1	0.1
Bangladesh	1.2	1.5	2.9	1.6	0.1	0.1	0.1
Bolivia	3.7	8.5	12.5	10.1	0.3	0.5	0.6
Bulgaria	1.0	1.3	1.3	3.0	0.4	0.4	0.3
Burkina Faso	0.4	0.5	0.9	0.5	0.1	0.1	0.0
Cambodia	0.4	0.5	0.7	0.7	0.2	0.2	0.1
Cameroon	1.2	1.9	3.3	1.4	0.0	0.0	0.0
Côte d'Ivoire	2.2	3.2	5.5	2.9	0.1	0.1	0.1
Ecuador	0.9	1.3	2.1	1.4	0.1	0.1	0.1
Ethiopia	1.1	1.4	2.5	1.2	-0.1	0.0	0.0
Georgia	1.6	2.2	3.5	2.5	0.1	0.2	0.2
Ghana	1.2	1.8	3.2	1.6	0.0	0.1	0.1
Guatemala	1.0	1.4	2.3	1.6	0.1	0.1	0.1
India	1.8	2.6	5.1	2.3	0.1	0.1	0.1
Iraq	1.1	1.2	1.8	1.4	0.1	0.1	0.1
Kyrgyzstan	1.3	1.2	1.4	2.1	0.5	0.3	0.2
Malawi	5.0	7.6	28.2	5.7	0.1	0.1	0.1
Mali	4.2	7.0	15.5	4.7	0.0	0.1	0.1
Mexico	4.8	8.3	13.7	8.9	0.3	0.5	0.4
Mongolia	0.6	0.7	0.9	1.6	0.1	0.1	0.1
Mozambique	5.2	9.0	16.5	9.4	0.7	0.8	0.7
Nepal	1.2	1.6	2.7	1.6	0.1	0.1	0.1
Nicaragua	0.6	1.0	1.4	1.3	0.2	0.2	0.1
Niger	0.7	0.9	1.6	0.9	0.0	0.0	0.0
Nigeria	6.0	9.7	27.7	7.4	0.4	0.4	0.3
Pakistan	0.7	0.9	1.4	1.0	0.3	0.2	0.1
Peru	0.8	1.2	1.8	1.4	0.1	0.1	0.1
Rwanda	3.4	5.6	11.5	4.3	0.3	0.3	0.2
Senegal	1.3	2.4	3.0	4.0	0.5	0.5	0.5
Serbia	1.1	1.3	1.8	1.9	0.3	0.2	0.2
Sierra Leone	0.5	0.8	0.9	1.3	0.3	0.3	0.2
Tanzania	0.8	1.2	1.8	1.1	0.0	0.0	0.0
Uganda	2.4	4.1	10.1	3.0	0.2	0.2	0.2
Vietnam	2.8	3.8	7.0	3.6	0.2	0.2	0.2
Average	*1.8*	*2.9*	*5.8*	*2.9*	*0.2*	*0.2*	*0.2*
SD	*1.6*	*2.8*	*7.2*	*2.6*	*0.2*	*0.2*	*0.2*
Min	*0.3*	*0.4*	*0.3*	*0.5*	*-0.1*	*0.0*	*0.0*
Max	*6.0*	*9.7*	*28.2*	*10.1*	*0.7*	*0.8*	*0.7*

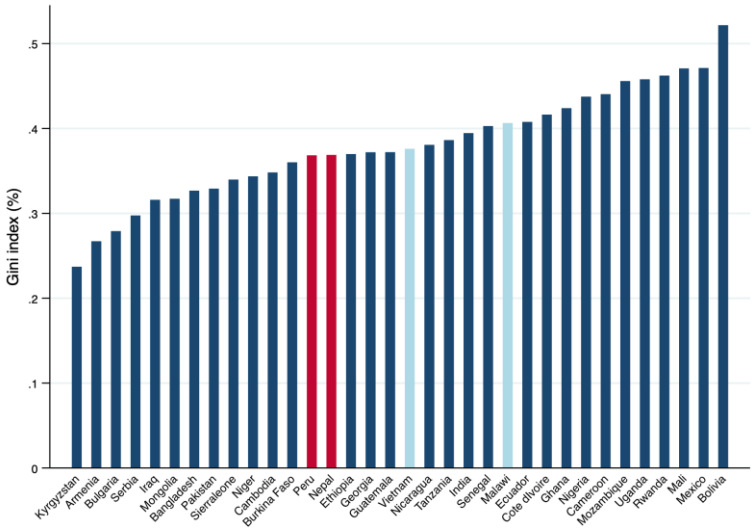
Note: MLD is for Mean Logarithmic Deviation; 'Atkinson' is for the Atkinson index, where the inequality aversion parameter epsilon is set equal to 2; H is for the poverty headcount ratio; PG is for the poverty gap; PG2 is for the poverty gap squared. Poverty indices are computed using a relative poverty line, equal to 60% of median per capita consumption in the original distribution.

Table 3 quantifies the sensitivity of an array of statistics of interest for welfare analysis. The first four columns refer to inequality measures: the Gini index (column 2), two members of the Generalized Entropy family of indices (Mean Log Deviation and Theil index, in columns 3 and 4), and the Atkinson index (with the inequality aversion parameter set equal to two, column 5). The remaining columns refers to Foster, Greer and Thorbecke (1984) poverty measures: the headcount ratio (H), the poverty gap index (PG), and the poverty gap squared index (PG2). The results show that the exclusion of outliers decreases estimated inequality according to all measures. The Gini index is the least sensitive among the inequality measures shown in Table 3, although before-after differences can be large enough to cast doubt on the validity of estimates obtained from the raw data: the Gini index for Nigeria, Mozambique, Malawi, and Mexico is bumped down by 5 to 6 percentage points, a salient difference by all metrics. The Theil index is the most sensitive measure among those we examined, showing differences as large as 28 percentage points. Overall, these results align with the theoretical analysis in Cowell and Flachaire (2007). Differences observed for three poverty measures – the poverty headcount, poverty gap, and poverty gap squared – are smaller across the board, in line with the theoretical findings available in the literature (Cowell and Victoria-Feser 1996b). However, our illustrative example uses relative poverty lines, which may affect results.

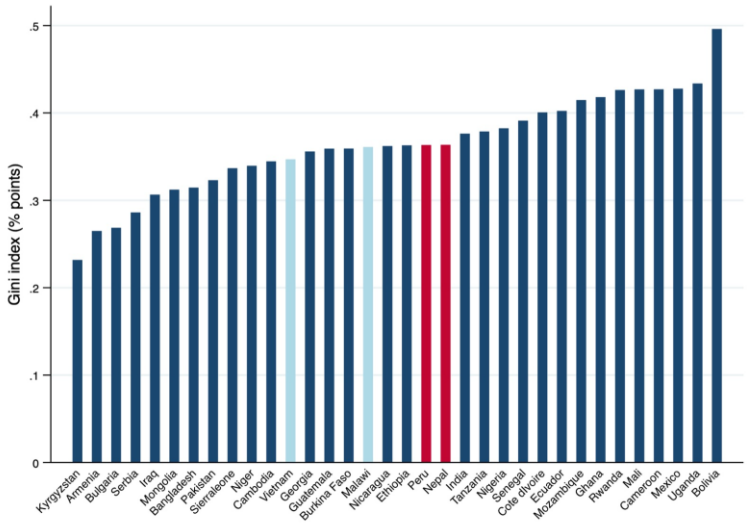
The empirical magnitude of the effects documented in Table 3 is consequential, and can be appreciated focusing on the case of Gini index (the more conservative scenario in our setting). Figure 6 (top panel) shows the point estimates of the Gini index calculated on the RuLIS 'raw data sets', that is on data as available from the data provider. Countries are ranked low (Kyrgyzstan) to high (Bolivia) accordingly. The bottom panel of Figure 6 shows the estimated Gini indices, for the same countries, after excluding outliers (as identified by `outdetect`). Even a cursory inspection of the two graphs bring to light a number of rerankings: countries move up or down the ladder as a consequence of how outliers are treated.

Figure 6 – Gini index (%), for selected countries

All observations



Without outliers



Source: our estimates based on RuLIS (2021).

Although the comparison between "raw" and "trimmed" statistics in Table 3 is meant as a sensitivity exercise, rather than an outlier treatment suggestion, clearly the behavior of the statistics of interest does depend on the choice of what to do with observations flagged as outliers. As an example, we consider winsorization (or censoring), which amounts to changing the value of each outlier to that of the nearest inlier (Tukey 1962: 18). In Table 4 we compare "raw" statistics with those obtained by winsorizing the per capita expenditure variable, rather than trimming it.

Table 4. Difference (in percentage points) between statistics computed before and after winsorizing the target variable

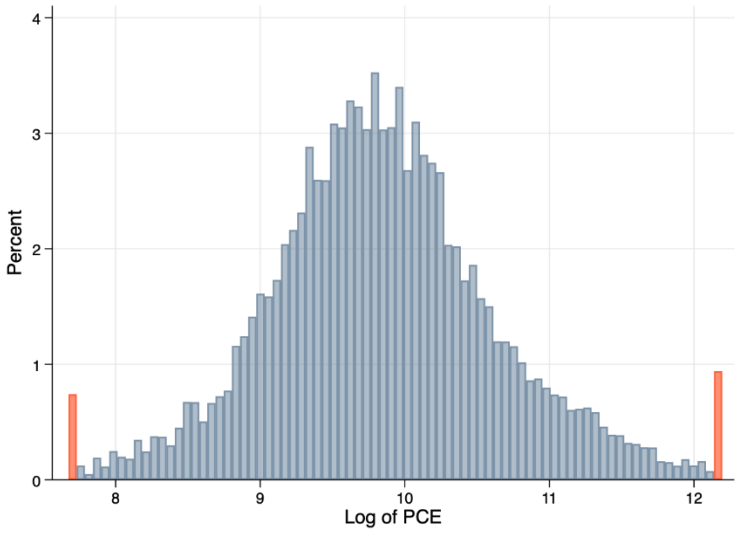
Country	Gini	MLD	Theil	Atkinson	H	PG	PG2
Armenia	0.0	0.1	0.0	0.2	0.0	0.0	0.0
Bangladesh	0.5	0.7	1.6	0.6	0.0	0.0	0.0
Bolivia	1.7	3.7	7.8	5.5	0.0	0.0	0.0
Bulgaria	0.3	0.5	0.5	1.4	0.0	0.1	0.1
Burkina Faso	0.0	0.1	0.2	0.1	0.0	0.0	0.0
Cambodia	0.1	0.1	0.2	0.2	0.0	0.0	0.0
Cameroon	0.4	0.7	1.6	0.4	0.0	0.0	0.0
Côte d'Ivoire	0.8	1.3	2.8	1.0	0.0	0.0	0.0
Ecuador	0.2	0.4	0.7	0.4	0.0	0.0	0.0
Ethiopia	0.4	0.6	1.3	0.5	0.0	0.0	0.0
Georgia	0.6	0.9	1.9	0.9	0.0	0.0	0.0
Ghana	0.4	0.7	1.5	0.6	0.0	0.0	0.0
Guatemala	0.2	0.4	0.7	0.6	0.0	0.0	0.0
India	0.7	1.2	2.8	0.9	0.0	0.0	0.0
Iraq	0.3	0.4	0.7	0.4	0.0	0.0	0.0
Kyrgyzstan	0.4	0.4	0.6	0.8	0.0	0.1	0.1
Malawi	3.9	6.2	26.2	4.2	0.0	0.0	0.0
Mali	2.5	4.3	11.3	2.6	0.0	0.0	0.0
Mexico	2.2	4.0	8.1	4.8	0.0	0.0	0.1
Mongolia	0.2	0.2	0.3	0.9	0.0	0.0	0.0
Mozambique	2.4	4.5	10.4	4.8	0.0	0.1	0.1
Nepal	0.4	0.5	1.1	0.4	0.0	0.0	0.0
Nicaragua	0.2	0.4	0.6	0.6	0.0	0.0	0.0
Niger	0.2	0.3	0.7	0.3	0.0	0.0	0.0
Nigeria	4.5	7.5	24.4	5.0	0.0	0.0	0.1
Pakistan	0.2	0.3	0.6	0.3	0.0	0.0	0.0
Peru	0.2	0.4	0.7	0.5	0.0	0.0	0.0
Rwanda	1.4	2.5	6.3	1.8	0.0	0.0	0.0
Senegal	0.5	0.9	1.5	1.8	0.0	0.0	0.1
Serbia	0.3	0.4	0.7	0.6	0.0	0.0	0.0
Sierra Leone	0.1	0.2	0.2	0.4	0.0	0.0	0.0
Tanzania	0.3	0.4	0.9	0.4	0.0	0.0	0.0
Uganda	1.5	2.7	8.0	1.8	0.0	0.0	0.0
Vietnam	1.4	2.0	4.4	1.6	0.0	0.0	0.0
Average	0.9	1.5	3.9	1.4	0.0	0.0	0.0
SD	1.1	1.9	6.3	1.6	0.0	0.0	0.0
Min	0.0	0.1	0.0	0.1	0.0	0.0	0.0
Max	4.5	7.5	26.2	5.5	0.0	0.1	0.1

Note: MLD is for Mean Logarithmic Deviation; 'Atkinson' is for the Atkinson index, where the inequality aversion parameter epsilon is set equal to 2; H is for the poverty headcount ratio; PG is for the poverty gap; PG2 is for the poverty gap squared. Poverty indices are computed using a relative poverty line, equal to 60% of median per capita consumption in the original distribution.

On average, differences between "raw" and "treated" results are about half as large when winsorizing (Table 4), as they are when trimming (Table 3), and patterns are consistent: the surveys where the impact of outliers is the largest remain the same (Nigeria, Malawi, Mexico…). This exercise is not intended to convey general messages but is a useful complement to the results presented in Table 3.

As to whether trimming or winsorization is to be preferred, no general recommendation can be found in the literature, nor any agreed-upon practices exist (van Kerm 2007; Turkiewicz, 2017; Hlasny 2020). When it comes to making a decision, it is worth reminding that winsorization produces artificial clusters of observations at the extremes of the distribution. The case of Mexico 2014 in Figure 7 illustrates: the histogram of the (log) of per capita expenditure shows two spikes (highlighted in red), created by winsorization. Depending on the statistics of interest to the analyst, the presence of these masses might be more or less influential: clearly, the choice between trimming and winsorizing has an impact on *cardinal* comparisons involving tail-sensitive poverty and inequality measures, which leads to recommendation that spatial and temporal poverty and inequality comparisons are carried out on a common method for treating outliers.

Figure 7 – Winsorized per capita expenditure, Mexico 2014



5 Concluding remarks

Despite the increase in quality of household surveys in recent years, most variables of interest to welfare analysts suffer from measurement error: the `outdetect` command focuses on outliers, that is, on values that are so large or so small that they seem "too far away" from the rest of the data. While 'outlier' is not necessarily a synonym for 'error', in practice their investigation often leads to the identification of gross errors; and even when outliers turn out to be legitimate values, their analysis is informative, and is integral to the initial stages of any data analysis. The literature provides three sets of results that motivate our interest in developing `outdetect` as a tool for detecting outliers and assessing their potential impact on poverty and inequality measures. First, the body of theoretical results that demonstrate the extent to which different inequality and poverty measures are sensitive to the presence of extreme values (Cowell and Victoria-Feser 1996a, 1996b; Victoria-Feser 2000; Cowell and Flachaire 2007, 2015; Cowell and van Kerm 2015). In general, welfare indices have been shown to be highly sensitive to extreme values in the tails of the distribution of income. Second, extreme values are omni-present in household surveys, both in low- and high-income countries (Hlasny and Verme 2018, Hlasny, Ceriani and Verme 2021). Third, while common experience and anecdotal evidence suggest that analysts always inspect their data, and often adjust them in some way, to protect their analyses from the impact of extreme values, these practices are rarely documented. Based on the examination of the documentation accompanying the release of some 200 official poverty and inequality estimates by national statistical offices, Mancini and Vecchi (2022) find that the overwhelming majority does not even mention whether outliers were dealt with, and how.

`outdetect` does not purport to solve the complex and long-standing problem of outlier identification and treatment once and for all. In fact, `outdetect` does not include options to treat extreme values, as no automatic procedures exist that can be recommended for that. Rather, the command should be thought of as a heuristic tool, much in the spirit of Hampel (1974) and Huber (1981), aimed at helping practitioners, particularly welfare analysts, with assessing the impact of extreme values in their calculations. Our effort in developing `outdetect` has focused on taking stock of both theoretical results and the practical constraints faced by empirical analysts. As a result, `outdetect` is simple to use, fast to execute even in the presence of relatively large samples, and rooted in the academic literature. The command is intended to greatly facilitate the documentation of choices made

at the “pre-analytical” stage, and the reproducibility of the analysis. This, in turn, has far-reaching implications for the consistency of welfare comparisons: when the methodology for identifying outliers differs, the consistency of both inter-temporal and international and intra-country geographic comparisons is at risk (Ravallion 1994; Atkinson 2019: ch. 4). Our preliminary empirical exploration, based on a selection of data sets drawn from the RuLIS project, provides a tentative assessment of the effect that lack of harmonization in dealing with outliers can have on welfare comparisons.

Appendix

Table A1. RuLIS data sets used in section 4

Country	Survey	Year	Households
Armenia	Integrated Living Conditions Survey	2013	5,184
Bangladesh	Household Income-Expenditure Survey	2010	12,240
Bolivia	Encuesta de los Hogares	2008	3,940
Bulgaria	Multitopic Household Survey	2007	4,300
Burkina Faso	Enquête Multisectorielle Continue	2014/15	10,800
Cambodia	Cambodia Socio-Economic Survey	2009	11,971
Cameroon	Fourth Cameroon Household Survey	2014	10,303
Côte d'Ivoire	Enquête Niveau de Vie des Menages 2008	2008	12,600
Ecuador	Encuesta sobre Condiciones de Vida	2014	28,970
Ethiopia	Ethiopia Socioeconomic Survey	2015/16	4,954
Georgia	Integrated Household Survey	2015	10,999
Ghana	Ghana Living Standards Survey	2012/13	16,772
Guatemala	Encuesta Nacional de Condiciones de Vida	2014	11,536
India	India Human Development Survey	2012	42,129
Iraq	The Iraq Household Socio-Economic Survey	2012	24,944
Kyrgyzstan	Integrated Sample Household Budget and Labour Survey	2013	5,013
Malawi	Integrated Household Survey	2017	12,447
Mali	Enquête Agricole de Conjoncture Integrée aux Conditions de Vie des Ménages	2014	3,804
Mexico	Encuesta Nacional de Ingresos y Gastos de los Hogares	2014	19,479
Mongolia	Socioeconomic Survey	2014	16,174
Mozambique	Inquérito sobre Orçamento Familiar	2009	10,832
Nepal	Nepal Living Standards Survey	2011	5,988
Nicaragua	Encuesta Nacional de Hogares sobre Medición de Nivel de Vida	2014	6,851
Niger	National Survey on Household Living Conditions and Agriculture	2014	3,617
Nigeria	General Household Survey	2012/13	4,738
Pakistan	Pakistan Social and Living Standards Measurement Survey	2013/14	17,989
Peru	Encuesta Nacional de Hogares	2015	32,188
Rwanda	Integrated Household Living Conditions Survey	2013/14	14,419
Senegal	Enquête de Suivi de la Pauvreté au Sénégal	2011	5,953
Serbia	Living Standards Measurement Survey	2007	5,557
Sierra Leone	Integrated Household Survey 2011	2011	6,727
Tanzania	National Panel Survey	2012/13	3,256
Uganda	The Uganda National Panel Survey	2013/14	3,352
Vietnam	Household Living Standards Survey	2010	9,399

Note: Information supplied by RuLIS technical documentation, http://www.fao.org/in-action/rural-livelihoods-dataset-rulis/technical-documentation/en/.

References

Atkinson, A. B. 2019. *Measuring poverty around the world*. Princeton University Press.

Barnett, V., and Lewis T. 1994. *Outliers in Statistical Data*. 3rd edition. J. Wiley and Sons.

Billor, N., Hadi, A. S. and Velleman P. F. 2000. "BACON: Blocked Adaptive Computationally Efficient Outlier Nominators. *Computational Statistics & Data Analysis*, 34: 279-298.

Box, G. E., and Cox, D. R. 1964. "An analysis of transformations." *Journal of the Royal Statistical Society*: Series B (Methodological), 26(2): 211-243.

Bruffaerts, C., Verardi, V. and Vermandele, C. 2014, "A generalized boxplot for skewed and heavy-tailed distributions." *Statistics and Probability Letters*, 95: 110-117.

Cowell, F. A. 2011. *Measuring Inequality*. Oxford University Press.

Cowell, F. A., and Flachaire, E. 2007. "Income distribution and inequality measurement: The problem of extreme values." *Journal of Econometrics*, 141(2): 1044-1072.

Cowell, F. A., and Flachaire, E. 2015. "Statistical methods for distributional analysis." In *Handbook of income distribution* (2: 359-465). Elsevier.

Cowell, F. A. and van Kerm, P. 2015. "Wealth Inequality: A Survey." *Journal of Economic Surveys*, 29(4): 671-710.

Cowell, F. A., and Victoria-Feser, M. P. 1996a. "Robustness Properties of Inequality Measures." *Econometrica*, 64(1): 77-101

Cowell, F. A., and Victoria-Feser, M. P. 1996b. "Poverty measurement with contaminated data: A robust approach." *European Economic Review*, 40(9): 1761-1771.

Croux, C. and Rousseeuw, P. J. 1992. "Tiem-efficient algorithms for two highly robust estimators of scale", *Computational Statistics*, 1: 411-428.

D'Agostino, R. B., Belanger, A., and D'Agostino R.B. Jr. (1990). "A Suggestion for Using Powerful and Informative Tests of Normality." *The American Statistician*, 44(4), 316–321.

Davies, L., and Gather, U. (1993). "The Identification of Multiple Outliers." *Journal of the American Statistical Association*, 88(423), 782–792.

Deaton, A. and Tarozzi, A. 2005. "Prices and poverty in India", in Deaton, A. and Kozel, V. (eds.), *The Great Indian Poverty Debate*. New Delhi, India: MacMillian: 381-411.

Dupriez, O. 2007. “Building a household consumption database for the calculation of poverty PPPs.” Technical note. The World Bank, Washington DC.

Filzmoser, P., Gussenbauer, J., and Templ, M. 2016. “Detecting outliers in household consumption survey data.” Vienna University of Technology.

Foster, J., Greer, J. and Thorbecke, E., 1984. “A class of decomposable poverty measures.” *Econometrica,* 761-766.

Friedline, T., Masa, R. D., and Chowa, G. A. 2015. “Transforming wealth: Using the inverse hyperbolic sine (IHS) and splines to predict youth’s math achievement.” *Social science research,* 49: 264-287.

Gather, U. and Becker, C. 1997. “Outlier Identification and Robust Methods”, in *Handbook of Statistics,* edited by Maddala, G. S. and Rao, C. R., vol. 15: 123-142.

Grubbs, F. E. 1969. “Procedures for detecting outlying observations in samples.” *Technometrics,* 11(1): 1-21.

Hampel, F. R. 1974. “The influence curve and its role in robust estimation.” *Journal of the American Statistical Association,* 69(346): 383–393.

Hampel, F. R., Ronchetti E. M., Rousseeuw P. J., and Stahel W. A. 1986. *Robust Statistics: The Approache Based in Influence Functions.* New York: John Wiley & Sons.

Hlasny, V. 2020. “Nonresponse Bias in Inequality Measurement: Cross-Country Analysis Using Luxembourg Income Study Surveys.” *Social Science Quarterly* 101, no. 2: 712–731.

Hlasny, V., Ceriani, L. and Verme, P. 2021. “Bottom Incomes and the Measurement of Poverty and Inequality.” *Review of Income and Wealth*.

Hlasny, V. and Verme, P. (2018) “Top Incomes and Inequality Measurement: A Comparative Analysis of Correction Methods Using the EU SILC Data.” *Econometrics,* 6(30): 1-21.

Huber, P. J. 2002. “John W. Tukey's Contributions to Robust Statistics.” *Annals of Statistics,* 30(6): 1640-1648.

Jann, B., V. Verardi and C. Vermandele. 2018. `robstat`: Stata module to estimate robust univariate statistics. Available from http://ideas.repec.org/c/boc/bocode/s458524.html.

Mancini, G. and Vecchi, G. 2022. *On the Construction of the Consumption Aggregate for Inequality and Poverty Analysis.* The World Bank: Washington DC.

Ravallion, M. 1994. *Poverty Comparisons*. London: Routledge.

Rousseeuw, P. J. and Croux, C . 1992. "Explicit Scale Estimators With High Breakdown Point", in *L_1 Statistical Analysis and Related Methods*, edited by Dodge, Y., Amesterdam, North Holland: 77-92.

Rousseeuw, P. J., and Croux, C. 1993. "Alternatives to the median absolute deviation." *Journal of the American Statistical Association*, 88(424): 1273-1283.

Rousseeuw, P., and Hubert, M. 2017. "Anomaly detection by robust statistics." *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 8(2): 1236, 1-14.

Shapiro, S. S., and Francia, R. S. 1972. "An approximate analysis of variance test for normality." *Journal of the American Statistical Association*, 67(337): 215-216.

Shapiro, S. S., and Wilk, M. B. 1965. "An analysis of variance test for normality (complete samples)." *Biometrika*, 52(3/4): 591-611.

Snedecor, G. W. and Cochran, W. G. 1989. *Statistical Methods*. Iowa State University Press, Ames.

Tukey, J. W. 1962. "The Future of Data Analysis," *The Annals of Mathematical Statistics* 33, no. 1: 1–67.

Turkiewicz, K. 2017. "Data trimming". In M. Allen (ed.), *The SAGE encyclopedia of communication research methods*, vol. 1: pp. 347-348. SAGE Publications.

Van Kerm, P. 2007. "Extreme Incomes and the Estimation of Poverty and Inequality Indicators from EU-SILC." *IRISS Working Paper Series* no. 2007-01.

Verardi, V. and Vermandele, C. 2018 "Univariate and Multivariate Outlier Identification for Skewed or Heavy-Tailed Distributions", *Stata Journal*, 18(3): 517-532.

Victoria-Feser, M. P. 2000. "Robust methods for the analysis of income distribution, inequality and poverty." *International Statistical Review*, 68(3): 277-293.

Weber, S. 2010. "bacon: An effective way to detect outliers in multivariate data using Stata (and Mata)." *Stata Journal*, 10(3): 331-338.

Yeo, I. K., and Johnson, R. A. 2000. "A new family of power transformations to improve normality or symmetry." *Biometrika*, 87(4): 954-959.