`outdetect`: Outlier detection for inequality and poverty analysis

Giulia Mancini (University of Sassari)
Federico Belotti (Tor Vergata University of Rome)
Giovanni Vecchi (Tor Vergata University of Rome)

19th Italian Stata Conference, Milan, September 25 2025

What is an outlier?

"An outlying observation, or 'outlier' is one that appears to deviate markedly from other members of the sample in which it occurs. (...) The observation may eventually be rejected as a result of the investigation, though not necessarily so."

Grubbs (1969)

The trouble with outliers: theory

- Outliers are omnipresent in real-world datasets, and they have the potential to bias most statistics of interest.
- Well-established literature on the effect of 'contaminants' on inequality and poverty estimates:
  - Cowell and Victoria-Feser (1996a) *Econometrica*
  - Cowell and Victoria-Feser (1996b) *EER*
  - Cowell and Flachaire (2007) *JoE*

The influence function

Cowell and Victoria-Feser (1996) *Econometrica*

F	CDF of ideal data, no contaminants
$I(F)$	"true" Gini index
$G = (1-\delta)F + \delta H,\ \ 0 \leq \delta \leq 1$	CDF of real-world data, with δ% contaminants
$I(G)$	"estimated" Gini index
$IF = \lim_{\delta \to 0} \frac{I(G) - I(F)}{\delta}$	Influence function (IF)

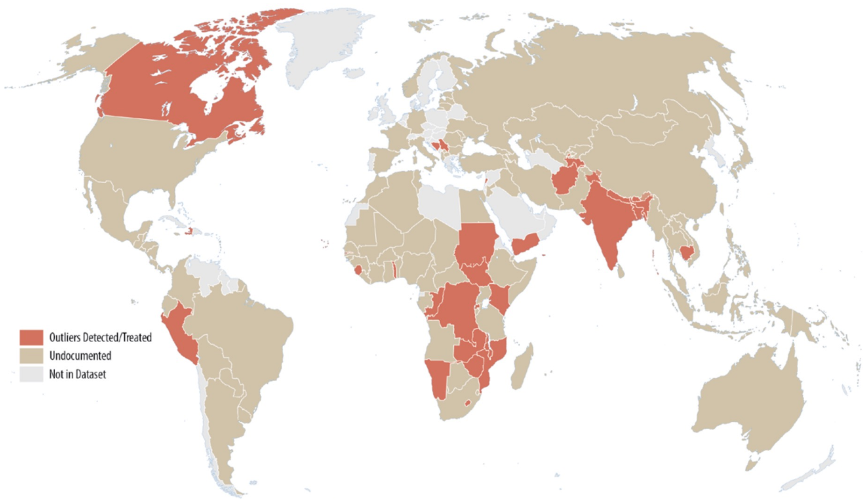
If the IF is unbounded, then $\mathrm{I(G)}$ can be infinitely biased for $\mathrm{I(F)}$.

The trouble with outliers: practice

- Using robust statistics is not an option in the context of welfare analysis: need for standard poverty and inequality indicators, and “clean” micro-data for public use.
- However, no shared conceptual framework for the detection and treatment of outliers.
- The issue often goes undocumented.

Official poverty and inequality reports

Mancini and Vecchi (2022)



Source: Mancini and Vecchi (2022) Appendix A

The `outdetect` command

1. Automates a popular outlier detection procedure, allowing for refinements that were previously cumbersome to implement in Stata;
2. Offers a variety of tools for assessing the impact of outliers on statistics of interest, *i.e.* sensitivity analysis.

- Emphasis on welfare analysis, but the procedure applies to any univariate, continuous, highly skewed distribution.

Outlier detection rule

- Consider target variable $X \sim f(x)$ (*e.g.* consumption expenditure per capita). Take a transformation $g(\cdot)$ such that $Y = g(X)$ is approximately Normal.
- Observation x_i is an oulier if:

$$z_i = \left| \frac{g(x_i) - \mu}{\sigma} \right| > z_\alpha$$

- μ and σ indicate the sample mean and standard deviation of Y, or any robust alternatives;
- z_α is a critical value of the distribution of (robust) z-scores, chosen by the analyst to identify a "low-probability" outlier region.

A popular application

- A popular version of this procedure is to log-transform the target variable, and flag observations that are more than two or three standard deviations from the mean.
- *E.g.* Deaton and Kozel (2005) contribution to the "great Indian poverty debate"
- `outdetect` offers a broader menu of normalizing tranformations and robust location and scale measures, to fit a variety of empirical applications.

Normalizing transformations

$$z_i = \left| \frac{\boxed{g(x_i)} - \mu}{\sigma} \right| > z_\alpha$$

Transformation	Formula
Square root	$y = \sqrt{x}$
Natural log	$y = \ln(x)$
Log (x+a)	$y = \log(x + a)$ with $a = \max[0, -\{\max(x) - 0.001\}]$
Box-Cox	$y = \begin{cases} \frac{x^\lambda - 1}{\lambda} & if\ \lambda \neq 0 \\ \ln(x) & if\ \lambda = 0 \end{cases}$
Yeo-Johnson	$y = \begin{cases} \frac{x^\lambda - 1}{\lambda} & if\ x \geq 0, \lambda \neq 0 \\ \ln(x + 1) & if\ x \geq 0, \lambda = 0 \\ \frac{-[(-x+1)^{(2-\lambda)} - 1]}{2 - \lambda} & if\ x < 0, \lambda \neq 2 \\ -\ln(-x + 1) & if\ x < 0, \lambda = 2 \end{cases}$
Inverse hyperbolic sine	$y = \ln\left(x + \sqrt{x^2 + 1}\right)$

Which one is best?

- Different transformations fit different needs (*e.g.* handling negative values of X)
- A choice can be made based on goodness-of-fit: the Pearson statistic P divided by its degrees of freedom df converges to 1 when the data approaches a Gaussian distribution (Peterson and Cavanaugh 2019)
- P/df can be used to rank transformations according to how successful they are in normalizing the data.

Measures of location and scale

$$z_i = \left|\frac{g(x_i) - \boxed{\mu}}{\boxed{\sigma}}\right| > z_\alpha$$

- Location: mean and median
- Scale: standard deviation and:

Estimator	Formula
IQR	$Q_3 - Q_1$
MAD	$1.4826 \operatorname{med}\lvert y_i - \operatorname{med}(y)\rvert$
S	$1.1926 \operatorname{med}_i[\operatorname{med}_j \lvert y_i - y_j\rvert]$
Q	$2.2219\,[\lvert y_i - y_j\rvert; i < j]_{(k)}$

The Median Absolute Deviation (MAD)

- MAD is defined as the median of the absolute deviations from the data's median.
- $MAD = 1.4826 \times med|x_i - med[x]|$
- The constant is an adjustment factor needed for using MAD as a consistent estimator of the standard deviation when the distribution is Normal
- MAD is probably the most popular robust estimator of the scale of a distribution.
- Rousseeuw and Croux (1993, JASA) suggest different options.

The S-estimator

- Rousseeuw and Croux (1993) propose to replace the MAD with the 'S estimator':
- $\mathrm{S} = \mathrm{c} \times \mathrm{med_i}\{\mathrm{med_j}|x_j - x_i|\}$, where $\mathrm{c} = 1.1926$ if the dn. is Normal
- Interpretation: for each observation i we compute the median $\mathrm{med_i}$ of the medians $\mathrm{med_j}$ of the absolute deviations $|x_j - x_i|$ for $\mathrm{j} = 1, \dots, \mathrm{n}$.

The Q-estimator

- Rousseeuw and Croux (1993) also propose the 'Q estimator':

  $\mathrm{Q} = \mathrm{d} \times \{|x_i - x_j|; i < j\}_{(k)}$, where $\mathrm{d} = 2.2219$ if the dn. is Normal

- Q is the k-th order statistic of the interpoint distances, with $\mathrm{k} \approx \binom{n}{2}/4$.

Recap

- Start from the original, untransformed variable X
- Transform X in order to obtain a variable Y 'as normally distributed as possible'
- Calculate a version of the z-score. Several candidates, including:

MAD	S	Q
$z_i^{MAD} = \left\lvert \frac{y_i - med[y_i]}{MAD} \right\rvert$	$z_i^{S} = \left\lvert \frac{y_i - med[y_i]}{S} \right\rvert$	$z_i^{Q} = \left\lvert \frac{y_i - med[y_i]}{Q} \right\rvert$

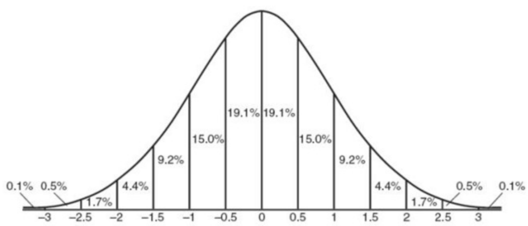
Which one is best?

- Debated.
- For practical purposes, the analyst may refer to the following ranking: $Q \succcurlyeq S \succ MAD \succ IQR$, where "$\succ$" stands for 'preferred to' and "$\succcurlyeq$" for 'weakly preferred to'

Outlier region threshold

$$z_i = \left| \frac{g(x_i) - \mu}{\sigma} \right| > \boxed{z_\alpha}$$

- The choice is not tied to any statistical requirement beyond the need to identify a "low-probability" outlier region.
- $z_\alpha = 3$ is customary; smaller and larger values (2.5, 3.5, or 4) are not uncommon.
- Higher values of z_α will shrink the outlier region and flag fewer outliers.



Basic syntax of `outdetect`

```
use https://raw.github.com/fbelotti/Stata/master/dta/hbs.dta, clear
svyset psu [pweight = weight]
outdetect pce
```

Variable `_out` is automatically generated:

0 = not an outlier

1 = bottom outlier

2 = top outlier

Basic output of `outdetect`

defaults

```
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	41	0.33	50.62
Top	40	0.32	49.38
Total	81	0.65	100.00

Statistics for raw and trimmed pce:

	Raw	Trimmed
Summary stats		
Mean	831.43	742.61
Median	575.71	575.89
SD	4714.51	598.31
CV(%)	567.03	80.57
IQR	509.46	505.03
Inequality		
Gini	0.4347	0.3662
GE(-1)	0.3539	0.2550
MLD	0.3309	0.2207
Theil	0.6726	0.2371
GE(2)	16.0751	0.3245
A(0.125)	0.0696	0.0290
A(1)	0.2817	0.1980
A(2)	0.4144	0.3378
p90/p10	5.1007	5.0361

Additional output: poverty measures

```
outdetect pce, pline(300000) replace
```

replace original _out variable

Statistics for raw and trimmed pce:

	Raw	Trimmed
Summary stats		
Mean	831.43	742.61
Median	575.71	575.89
SD	4714.51	598.31
CV(%)	567.03	80.57
IQR	509.46	505.03
Inequality		
Gini	0.4347	0.3662
GE(-1)	0.3539	0.2550
MLD	0.3309	0.2207
Theil	0.6726	0.2371
GE(2)	16.0751	0.3245
A(0.125)	0.0696	0.0290
A(1)	0.2817	0.1980
A(2)	0.4144	0.3378
p90/p10	5.1007	5.0361
Poverty		
H	0.1352	0.1331
PG	0.0332	0.0314
PG2	0.0119	0.0106

Customize detection procedure

```
outdetect pce, norm(bcox) zscore(mean std) alpha(2) out(top) replace
```

choose normalizing transformation

choose z-score

choose outlier threshold

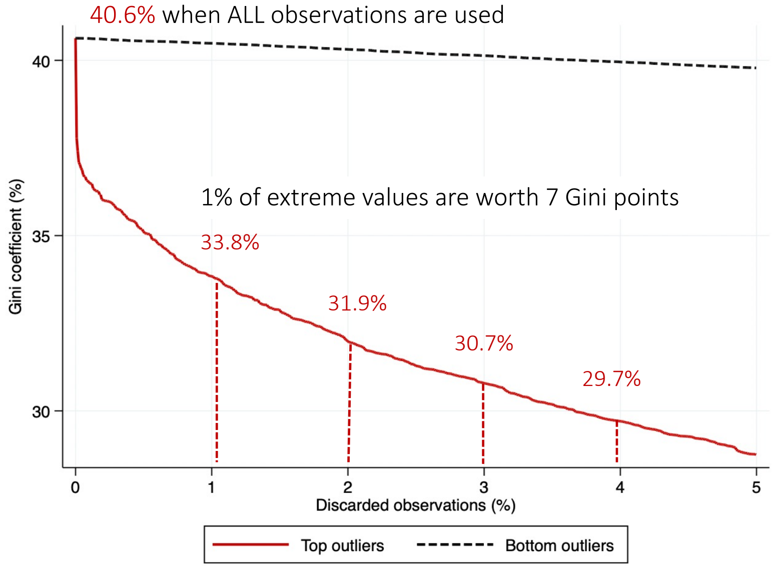
focus on bottom or top outliers

```
outdetect pce, bestnormalize replace
```

select the 'best' normalizing transformation according to P/df

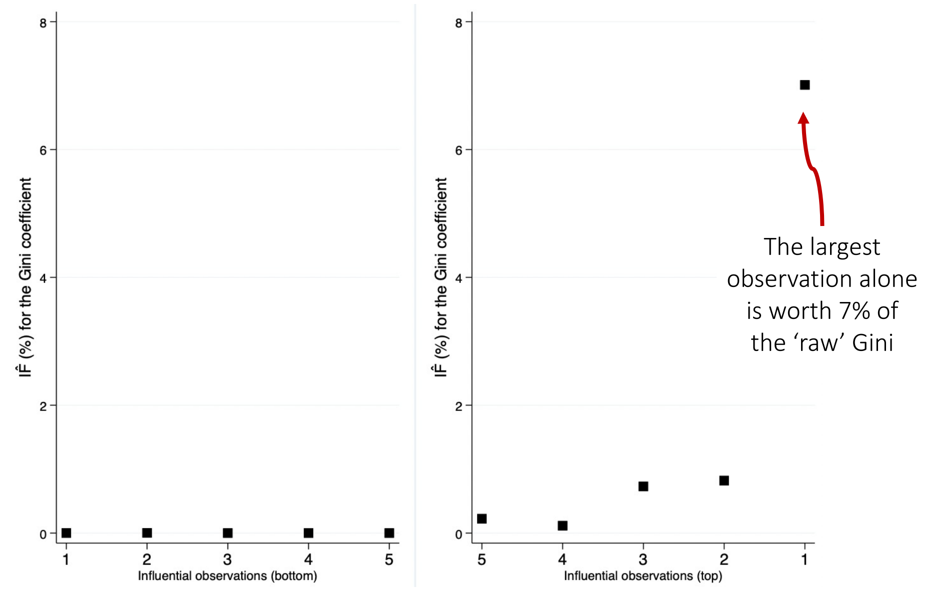
Diagnostic graphs: Incremental Trimming Curve

- The ITC plots the value of any statistic of interest calculated on the distribution of X after sorting the values of X and excluding the first i cumulated observations.
- The statistic of interest and the number of observations discarded are customizable.
- `outdetect pce, graph(itc)`



Diagnostic graphs: Influence Function Curve

- Proposed by Cowell and Flachaire (2007)
- The IFC plots the percentage difference between the ‘actual’ value of the statistic of interest, and the value calculated after excluding the i-th observation of X.
- The statistic of interest and the number of observations discarded are customizable.
- `outdetect pce, graph(ifc)`



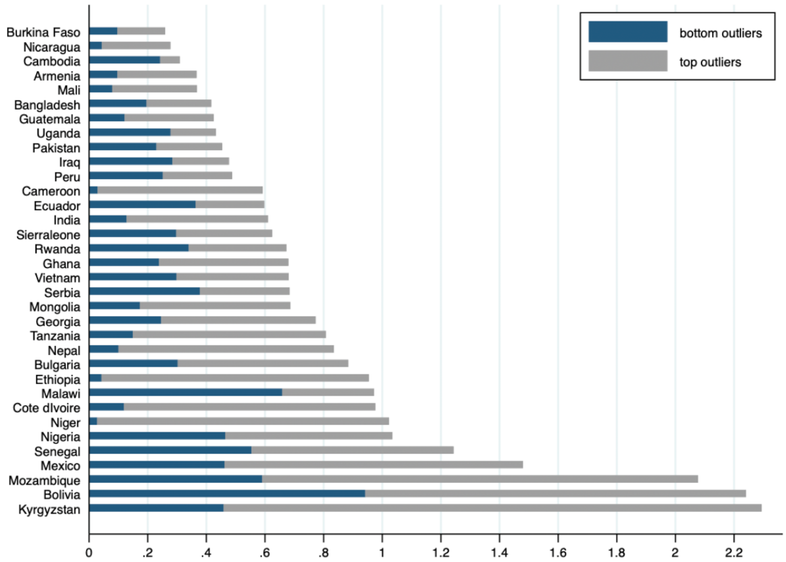
Other options and goodies

- Generate a new variable containing the normalized target variable (`normvar`)
- Draw diagnostic plots of normalized variable (`qqplot`, `qqpareto`, `zipf`)
- Export statistics in Excel (`excel (`*`filename`*`)`)
- Recompute survey weights after exclusion of outliers (`reweight`)

`outdetect` in action

- Rural Livelihoods Information System (RuLIS): joint initiative of the Food and Agriculture Organization (FAO), the World Bank, and the International Fund for Agricultural Development (IFAD).
- We use data from 34 countries, apply `outdetect`'s default detection rule to per capita consumption, and assess the sensitivity of poverty and inequality measures.

Incidence of outliers (% of N)

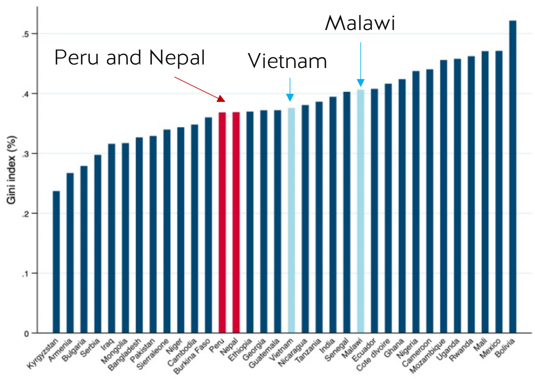


Difference
(in percentage points)
before-after exclusion of
outliers

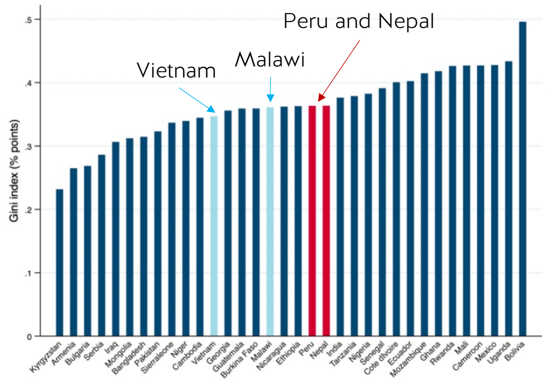
Country	Gini	MLD	Theil	Atkinson	H	PG	PG2
Armenia	0.3	0.4	0.3	0.8	0.2	0.1	0.1
Bangladesh	1.2	1.5	2.9	1.6	0.1	0.1	0.1
Bolivia	3.7	8.5	12.5	10.1	0.3	0.5	0.6
Bulgaria	1.0	1.3	1.3	3.0	0.4	0.4	0.3
Burkina Faso	0.4	0.5	0.9	0.5	0.1	0.1	0.0
			(...)				
Rwanda	3.4	5.6	11.5	4.3	0.3	0.3	0.2
Senegal	1.3	2.4	3.0	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*

Gini index for selected countries

All observations



Without outliers



The Stata Journal (2024)
24, Number 4, pp. 630–665 DOI: 10.1177/1536867X241297918

Outlier detection for inequality and poverty analysis

Federico Belotti
Tor Vergata University of Rome
Rome, Italy
federico.belotti@uniroma2.it

Giulia Mancini
University of Sassari
Sassari, Italy
gmancini@uniss.it

Giovanni Vecchi
Tor Vergata University of Rome
Rome, Italy
giovanni.vecchi@uniroma2.it

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. In this article, we discuss a simple univariate detection procedure to flag outliers. We present `outdetect`, a command that implements the procedure and provides useful diagnostic tools. The output of `outdetect` compares statistics obtained before and after the exclusion of outliers, with a focus on inequality and poverty measures. Finally, we carry 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 and provide a sense of the influence of extreme values on poverty and inequality estimates.

Keywords: st0759, outdetect, outliers, extreme values, inequality, poverty, incremental trimming curve

Thank you for your attention!