

# Outlier detection and treatment

LECTURE 12

1

---

---

---

---

---

---

---

---

Today is mainly about outliers

- 1) **Definitions**  
What do we mean by an outlier, exactly?
- 2) **Motivation**  
Do outliers really matter?
- 3) **Detection**  
How to detect outliers?
- 4) **Treatment**  
How to deal with outliers?

2

---

---

---

---

---

---

---

---

Important premise

- Suggestions shared in this lecture are not a substitute for the protocols that NSOs have in place to ensure data quality
- They are meant to offer **further safeguards** once **"routine" edits** have been completed
- Useful to **analysts**, as well as data producers

3

---

---

---

---

---

---

---

---

# Definitions

4

---

---

---

---

---

---

---

---

# What is an outlier?

An outlier is an observation "that appears to deviate markedly from other members of the sample in which it occurs" (Grubbs, 1969)



5

---

---

---

---

---

---

---

---

# What is an outlier?

- Note: we focus on **univariate** outliers, those found when looking at a distribution of values in a single dimension (e.g. income).
- We use Venice to illustrate



6

---

---

---

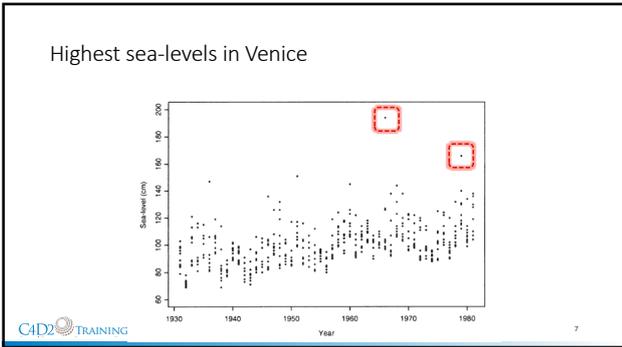
---

---

---

---

---



7

---

---

---

---

---

---

---

---

### What causes outliers?

- **Human errors**, e.g. data entry errors
- **Instrument errors**, e.g. measurement errors
- **Data processing errors**, e.g. data manipulation
- **Sampling errors**, e.g. extracting data from wrong sources
- **Not an error**, the value is extreme, just a 'novelty' in the data

C4D2 TRAINING

8

---

---

---

---

---

---

---

---

### A dilemma

- Outliers can be **genuine** values
- The trade-off is between the loss of **accuracy** if we throw away "good" observations, and the **bias** of our estimates if we keep "bad" ones
- The challenge is twofold:
  1. to figure out whether an extreme value is good (genuine) or bad (error)
  2. to assess its impact on the statistics of interest

C4D2 TRAINING

9

---

---

---

---

---

---

---

---

Do outliers matter?



10

10

---

---

---

---

---

---

---

---

Theory first

- Three papers:
  - I. 1996a  
Frank Cowell and Maria-Pia Victoria-Feser
  - II. 2007  
Frank Cowell and Emmanuel Flachaire (\*)
  - III. 1996b  
Frank Cowell and Maria-Pia Victoria-Feser



11

---

---

---

---

---

---

---

---

Outliers and inequality measures – I  
Cowell and Victoria-Feser (1996a)

Economica, Vol. 64, No. 1 (January, 1996), 77-100

ROBUSTNESS PROPERTIES OF INEQUALITY MEASURES<sup>1</sup>

BY FRANK A. COWELL AND MARIA-PIA VICTORIA-FESER<sup>2</sup>

Inequality measures are often used to summarise information about empirical income distributions. However the resulting picture of the distribution and of changes in the distribution can be severely distorted if the data are contaminated. The robustness of the measures will in general depend upon the sampling properties of the inequality measure. We investigate this issue theoretically using a technique based on the influence function, and illustrate the magnitude of its effect using a simulation. We consider both point and regression estimates with the simple and robust estimates using a parametric model; in the latter case we demonstrate the application of a robust estimation procedure. We apply our results to two micro-data exercises.

Keywords: Inequality, contaminated data, influence function, parametric estimation, normal distribution.

- This is a beautiful paper
- Explains why outliers (contaminants) are a serious threat to most inequality measures.
- “if the mean has to be estimated from the sample then all scale independent or translation independent and decomposable measures have an unbounded influence function” (p. 89)
- An unbounded IF is a catastrophe.



12

---

---

---

---

---

---

---

---

### Why an unbounded IF is a catastrophe



- The IF is a measure of the bias of an estimator due to the presence of extreme values.
- An unbounded IF means that the bias can be infinitely large.
- If the bias of inequality estimators can be infinitely large, outliers are a priority for both data producers and data users.

15

---

---

---

---

---

---

---

---

### In practice

Hlasny and Verme (2018: 191)

- Many researchers routinely trim outliers or problematic observations or apply top coding with little consideration of the implications for the measurement of inequality
- One example to illustrate

16

---

---

---

---

---

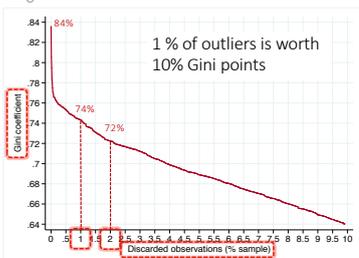
---

---

---

### Sensitivity of the Gini index to extreme values

iterative trimming



17

---

---

---

---

---

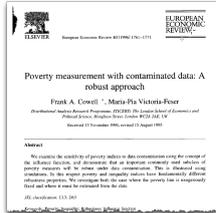
---

---

---

## Outliers and poverty measures

Cowell and Victoria-Feser (1996b)



- Explains why **outliers** only **rarely** are a serious **threat** to most poverty measures.
- Poverty measures are not sensitive to the values (real or contaminated) of the incomes of the rich

20

---

---

---

---

---

---

---

---

## Recap

- The answer to the question on whether outliers matter **depends** on the statistic of interest
- **Inequality**: both theory (unbounded IF) and practice (incremental truncation) suggest that they matter (tremendously). Not taking this issue into proper account puts inequality comparisons at risk.
- **Poverty**: not so much

21

---

---

---

---

---

---

---

---

## How to detect outliers?

22

---

---

---

---

---

---

---

---

## Visual inspection

(Deaton and Tarozzi 2005)

- “Our procedures are part **graphical**, and part **automatic**. For each commodity, we draw histograms and one-way plots of the logarithms of the unit values, using each to detect the presence of gross outliers for further investigations. [...] [Automatic method] **does not remove the need** for the graphical inspection”

23

---

---

---

---

---

---

---

---

## Visual inspection

Malawi IHS3, Cassava tuber expenditure

**Malawi - Integrated Household Panel Survey 2010-2013 (Short-Term Panel, 204 EAs)**

Reference ID	MHW_2010-2013_IHPS_v01_M	Created on	Apr 21, 2010
Year	2010 - 2013	Last modified	Dec 13, 2017
Country	Malawi		
Producer(s)	National Statistical Office - Government of Malawi		
Sponsor(s)	Government of Malawi - SIDA/SIDA - Funded the study The World Bank - WFP - Funded the study Hilary Clinton Challenge Corporation - HCCC - Funded the study Sida S4U - S4U - Funded the study German Development Cooperation - GIZ - Funded the study		
Collection(s)	Living Standards Measurement Study (LSMS) Central		
Metadata	Documentation in PDF		
Related Materials	Study Description	Data Dictionary	Get Microdata

24

---

---

---

---

---

---

---

---

## Visual inspection

Malawi IHS3, Cassava tuber expenditure

### MODULE G: FOOD CONSUMPTION OVER PAST ONE WEEK

DATA ENTRY LINE NUMBER	Question	G01	G02	G05
	Over the past one week (7 days), did you or others in your household consume any [...]?  INCLUDE FOOD BOTH EATEN COMMUNALLY IN THE HOUSEHOLD AND THAT EATEN SEPARATELY BY INDIVIDUAL HOUSEHOLD MEMBERS.	YES . . 1 NO . . . 2 >> NEXT ITEM	ITEM CODE	How much did you spend?
19	Roots, Tubers, and Plantains			
20	Cassava tubers		201	

25

---

---

---

---

---

---

---

---

### Visual inspection

Example 1: look at descriptive statistics

```
. sum hh_g05 if hh_g02==201.d
```

How much did you spend?

Percentiles	Smallest		
1%	5	0	
5%	20	0	
10%	20	0	Obs 673
25%	50	0	Sum of Wgt. 673
50%	75	Mean	94.95097
		Std. Dev.	106.2379
75%	100	400	
90%	200	400	Variance 11286.5
95%	220	1050	Skewness 10.0151
99%	350	2000	Kurtosis 164.7054

26

---

---

---

---

---

---

---

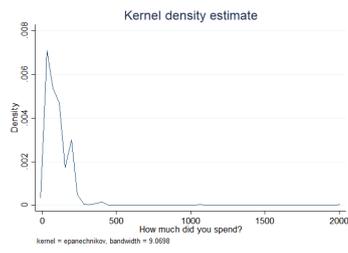
---

---

---

### Visual inspection

Example 2: graph the distribution of the data



27

---

---

---

---

---

---

---

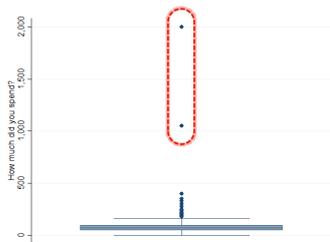
---

---

---

### Visual inspection

Example 3: use graphical diagnostic tools, e.g. the boxplot graph



28

---

---

---

---

---

---

---

---

---

---

## Statistical methods

- The literature is rich with methods to identify outliers; in practice, most methods used in empirical work hinge on the underlying **distribution of the data**.
- The idea is simple:
  - **transform** the variable to induce **normality**
  - **set thresholds** to identify extreme values

29

---

---

---

---

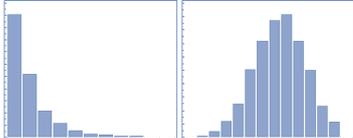
---

---

---

## Transform the variable to induce normality

- The easiest transformation relies on **taking the logarithm** of the variable of interest
- The log “squeezes” large values more, so that skewed distributions become more symmetrical and closer to a Normal distribution.



30

---

---

---

---

---

---

---

## Set a threshold

- We must specify a **threshold** for deciding whether each observation is ‘too extreme’ (outlier or not?)
- Common ‘thumb-rule’ **thresholds** : an observation is considered an outlier if it is more than **2.5, 3, 3.5 standard deviations far from the mean** of the distribution
- In formulas:  $x$  is an outlier if  $x > \bar{x} + z_\alpha s$   
where  $z_\alpha$  equals, say, 2.5.
- We can express the same criterion as  $\frac{x - \bar{x}}{s} > z_\alpha$   
where the left-hand side is called a **z-score** (a variable with mean = 0 and var = 1)

31

---

---

---

---

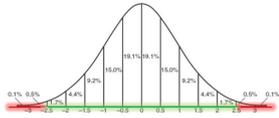
---

---

---

Why 2.5, 3, 3.5, or any other number?

- Under the assumption of normality:



$z_\alpha = 2.5$  implies that outliers are in the region where  $\alpha = 0.5$  percent of observations normally are.

32

---

---

---

---

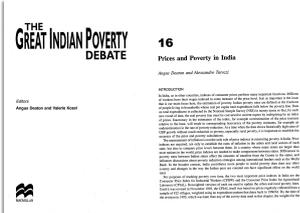
---

---

---

---

Deaton and Tarozi (2005)



In the case of India, D&T (2005) flagged as outliers prices whose logarithms exceeded the mean of logarithms by more than 2.5 standard deviations:

$$\frac{\ln(x) - E[\ln(x)]}{sd[\ln(x)]} > 2.5$$

33

---

---

---

---

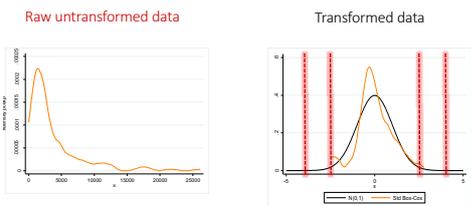
---

---

---

---

Transformation and thresholds



34

---

---

---

---

---

---

---

---

### Two questions

- 1) How good is such an approach?
- 2) What to do after flagging outliers?

35

---

---

---

---

---

---

---

### How good is such an approach?

- Log-transformation is very basic – e.g. how to deal with negative values?
- Not recommended when the log-distribution cannot be assumed to be Normal
- Why should we set the threshold using the **mean** and **standard deviation**, which are sensitive to extreme values, if this is exactly what we are worried about?

$$\frac{\ln(x) - E[\ln(x)]}{sd[\ln(x)]} > 2.5$$

- We can do better

36

---

---

---

---

---

---

---

### A popular strategy

robustification

- While there is no agreement on the best method, a common solution is to use **robust measures of scale** and **location** to set the threshold for flagging outliers
- the idea is to replace the **sample average**  $\bar{x}$  with a robust estimator (e.g. the **median**), and the **standard deviation**  $s$  with a robust estimator. A popular option is the **median absolute deviation (MAD)**.

39

---

---

---

---

---

---

---

Robustification of the z-score  
The median absolute deviation (MAD)

$$z_h = \frac{x_h - \bar{x}}{s}$$

$$z_h = \frac{x_h - \text{med}[x_h]}{MAD}$$

$$MAD = b \times \text{med}[x - \text{med}[x]]$$

$$b = 1.4826$$

if the distribution is Gaussian



40

---

---

---

---

---

---

---

---

We can do even better  
Rousseeuw and Croux (1993, JASA)

Alternatives to the Median Absolute Deviation

Peter J. Rousseeuw and Christophe Croux\*

In robust estimation one frequently needs an initial or auxiliary estimate of scale. For this one usually takes the median absolute deviation  $MAD_n = 1.4826 \text{med}\{|x_i - \text{med}(x_i)|\}$  because it has a simple explicit formula, needs little computation time, and is very robust as witnessed by its bounded influence function and its 50% breakdown point. But there is still room for improvement in two areas: the fact that MAD is aimed at symmetric distributions and its low (37%) Gaussian efficiency. In this article we set out to construct explicit and 50% breakdown scale estimators that are more efficient. We consider the estimator  $S_n = 1.1926 \text{med}\{\text{med}\{|x_i - x_j|\}\}$  and the estimator  $Q_n$  given by the 23 quantile of the distances  $\{|x_i - x_j|; i < j\}$ . Note that  $S_n$  and  $Q_n$  do not need any location estimate. Both  $S_n$  and  $Q_n$  can be computed using  $O(n \log n)$  time and  $O(n)$  storage. The Gaussian efficiency of  $S_n$  is 58%, whereas  $Q_n$  attains 82%. We study  $S_n$  and  $Q_n$  by means of their influence functions, their bias curves (or implusions as well as explosions), and their finite-sample performance. Their behavior is also compared at non-Gaussian models, including the negative exponential model where  $S_n$  has a lower gross error sensitivity than the MAD.

KEY WORDS: Bias curve; Breakdown point; Influence function; Robustness; Scale estimation.



41

---

---

---

---

---

---

---

---

Rousseeuw and Croux (1993)

- Rousseeuw and Croux (1993) propose to substitute the MAD with a different estimator:

$$S = c \times \text{med}\{\text{med}\{|x_j - x_i|\}\}$$

- For each  $j$  we compute the median of  $|x_j - x_i|$  ( $i = 1, \dots, n$ ). This yields  $n$  numbers, the median of which gives our final estimate  $S$ .

$$z_h = \frac{x_h - \text{med}[x_h]}{S}$$

$c = 1.1926$  at the Gaussian model.



42

---

---

---

---

---

---

---

---

## Recap

- We can do better than “take the log and run”:
  1. Using **transformations other than the log** to normalize the variable of interest often gives better results
  2. **Robustifying the z-score** is a better practice.
- Belotti et al. (2020): [outdetect.ado](https://outdetect.ado) helps with both things

43

---

---

---

---

---

---

---

---

## Take the log and run vs. robust z-scores

Countries	Year	Outliers (%)					
		log-transformation			robust z-scores		
cutoff = 3		overall	left	right	overall	left	right
		(1)	(2)	(3)	(4)	(5)	(6)
Malawi	2017	0.75	0.14	0.61	0.30	0.22	0.08
Nigeria	2012	1.35	0.11	1.24	0.72	0.32	0.40
India	2012	1.39	0.03	1.36	0.62	0.13	0.49
Pakistan	2014	1.58	0.02	1.56	0.39	0.21	0.18
Guatemala	2014	1.14	0.06	1.08	0.61	0.15	0.46
Peru	2015	0.36	0.09	0.27	0.28	0.16	0.12
Armenia	2013	0.91	0.08	0.83	0.68	0.17	0.51
Georgia	2015	0.75	0.25	0.50	0.73	0.32	0.41

44

44

---

---

---

---

---

---

---

---

## Impact of outliers on the Gini index

Countries	Year	Gini index		
		Raw	Trimmed (log)	Trimmed (best)
cutoff = 3		(7)	(8)	(9)
Malawi	2017	40.6	34.8	36.6
Nigeria	2012	43.7	36.7	38.2
India	2012	39.5	36.2	37.6
Pakistan	2014	32.9	30.0	32.3
Guatemala	2014	37.2	34.7	35.9
Peru	2015	36.8	36.0	36.3
Armenia	2013	28.9	26.7	26.9
Georgia	2015	37.1	35.4	35.6

45

45

---

---

---

---

---

---

---

---

# How to deal with outliers?

(in one slide)



46

---

---

---

---

---

---

---

---

46

## Treatment of outliers

Three main methods for dealing with outliers, apart from removing them from the dataset:

- 1) **reducing the weights** of outliers (trimming weight)
- 2) **changing the values** of outliers (Winsorisation, trimming, imputation – for instance via quantile regression)
- 3) **using robust estimation techniques** (M-estimation).

- Documentation, transparency and reproducibility



47

---

---

---

---

---

---

---

---

47



## Lessons learned

- Outliers can be **genuine** observations... be gentle to the data and document each and every step of the data processing
- As far as inequality is concerned, outliers are the worst enemy (**unbounded IF**)
- Outlier detection:
  - Go beyond the “**take the log and run**” strategy. It works well only if you can describe the data with a Gaussian distribution. Typically, however, distributions are skewed.
  - Instead, use the “**best-normalization**” available, and **robustify the z-score**.
- **Outlier treatment**: it depends. Quantile regression is a good candidate.



48

---

---

---

---

---

---

---

---

48

## References

### Required readings

Barnett, V., & Lewis, T. (1994). Outliers in Statistical Data. 3rd edition. J. Wiley & Sons (Chapter 1 & 2)

### Suggested readings

Alvarez, E., Garcia-Fernández, R. M., Blanco-Encinosa, F. J., & Muñoz, J. F. (2014). The effect of outliers on the economic and social survey on income and living conditions. *World Acad. Sci., Eng. Technol., Int. J. Soc., Behav., Educ., Econ., Bus. Ind. Eng.*, 8, 3276-3280.

Baloni, F., & Vecchi, G. (2019). Take the Log and Run: Outliers and Welfare Measurement, mimeo.

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

Cowell, F., & Victoria-Feser, M. (1996). Robustness Properties of Inequality Measures. *Econometrica*, 64(1), 77-101.

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

Deshmukh, A., & Tansik, A. (2005). "Prices and Poverty in India." *The Great Indian Poverty Debate*. New Delhi : MacMillan.

Dupriez, O. (2007). Building a household consumption database for the calculation of poverty PPPs. Technical note. Available at: <http://go.worldbank.org/4YGT5RGT0>.

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

Hlasny, V., & Verme, P. (2018). Top Incomes and Inequality Measurement: A Comparative Analysis of Correction Methods Using the EU SILC Data. *Econometrics*, 6(2), 30.

Mancini, G., & Vecchi, G. (2019). On the Construction of a Welfare Indicator for Inequality and Poverty Analysis, mimeo.

OECD (2013). OECD Guidelines for Micro Statistics on Household Wealth

Rousseleux, P. J., & Croux, C. (1993). Alternatives to the median absolute deviation. *Journal of the American Statistical association*, 88(424), 1273-1283.

---

---

---

---

---

---

---

---

Thank you for your attention

---

---

---

---

---

---

---

---

Homework

---

---

---

---

---

---

---

---



### Exercise 3 – Inequality measures

OECD (2013)

Table 7.3. Effect of the treatment of outliers on summary measures of wealth inequality in the United States, 2007

	Raw	Shave top and bottom 1%	Shave top 1% and bottom 0.5%
Mean	556 846	378 215	559 361
Median	120 780	120 780	123 800
Gini	0.82	0.74	0.81
$\frac{1}{2}CV^2$	18.1	2.4	14.6
P90/P10	30 000	3 369	3 061
P75/P25	26.3	24.5	24.3
P90/P50	7.6	7.0	7.4
<i>n</i>	4 418	3 698	4 359

Source: 2007 Survey of Consumer Finances.

---

---

---

---

---

---

---

---