Robust measures of income and wealth inequality

Giovanni Vecchi
U. Rome “Tor Vergata”

C4D2 – Perugia
December 10-14, 2018

Two questions

1) How to produce robust estimates of wealth (income) inequality?

   robust = resilient to data flaws

2) Is there a 'best international practice' to deal with data flaws?

   this improves comparability and international harmonization

Two introductory thoughts

- The general public often focuses on levels
  - How large is households' private wealth?
  - How high is wealth inequality?
- Most important questions, however, typically imply comparisons
  - Has net worth increased during the last year?
  - Has wealth inequality become more concentrated?
- Despite the many data repositories available, comparing income and wealth over time and across space is no easy task

Harmonized international datasets

World Bank World Development Indicators (WDI)

UNU-WIDER World Income Inequality Database (WIID)

Luxembourg Income Study

Luxembourg Wealth Study

FAO RuLIS

OECD Social and Welfare Statistics

European Central Bank
Household Finance and Consumption Survey (HFCS)

Harmonized international datasets - links

http://databank.worldbank.org/data/reports.aspx?source=world-development-indicators

https://www.wider.unu.edu/project/wiid-world-income-inequality-database

http://www.lisdatacenter.org/our-data/lis-database/

http://www.lisdatacenter.org/our-data/lws-database/

(coming soon)

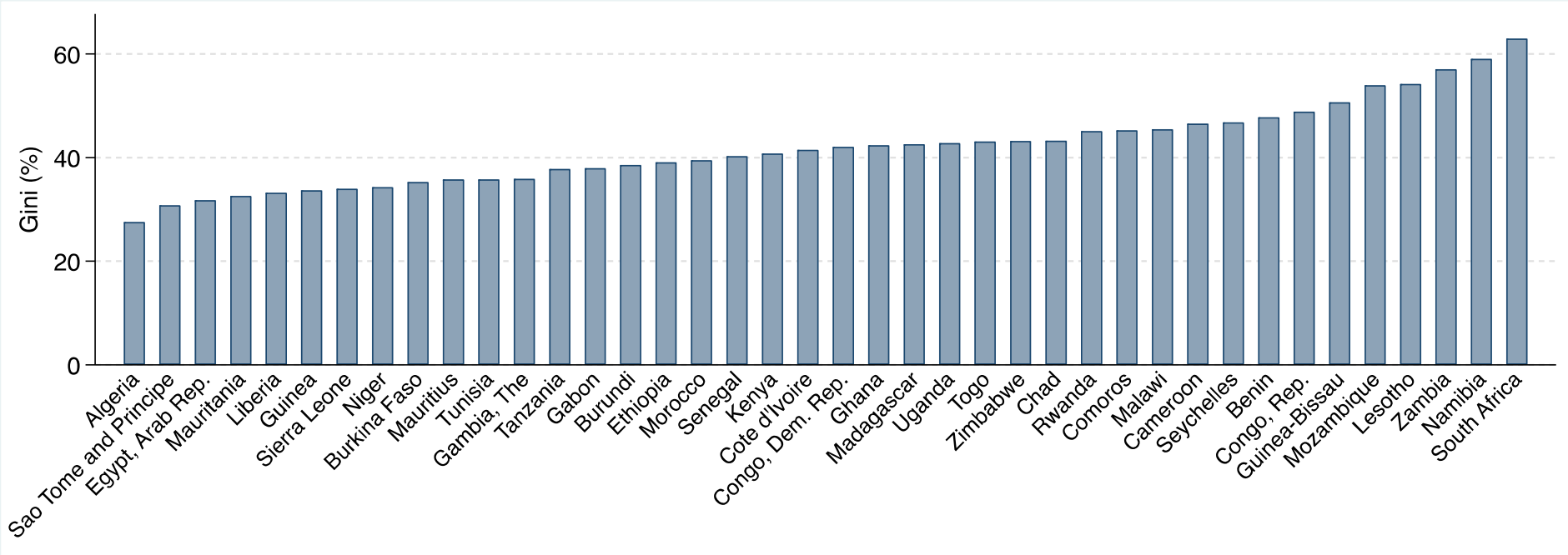
https://www.oecd-ilibrary.org/social-issues-migration-health/data/oecd-social-and-welfare-statistics_socwel-data-en

https://www.ecb.europa.eu/pub/economic-research/research-networks/html/researcher_hfcn.en.html

International comparisons

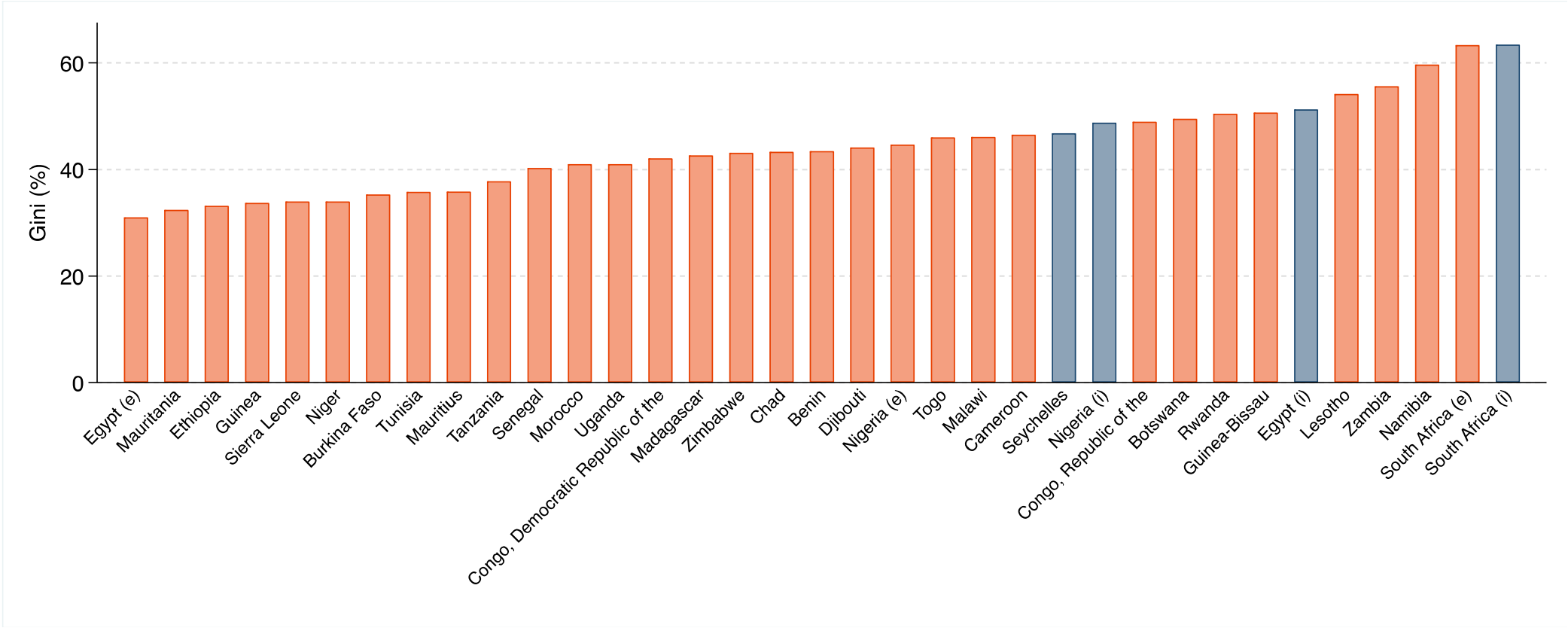
How robust is the ranking?

World Bank, WDIs – Gini index



UNU-WIDER, WIID

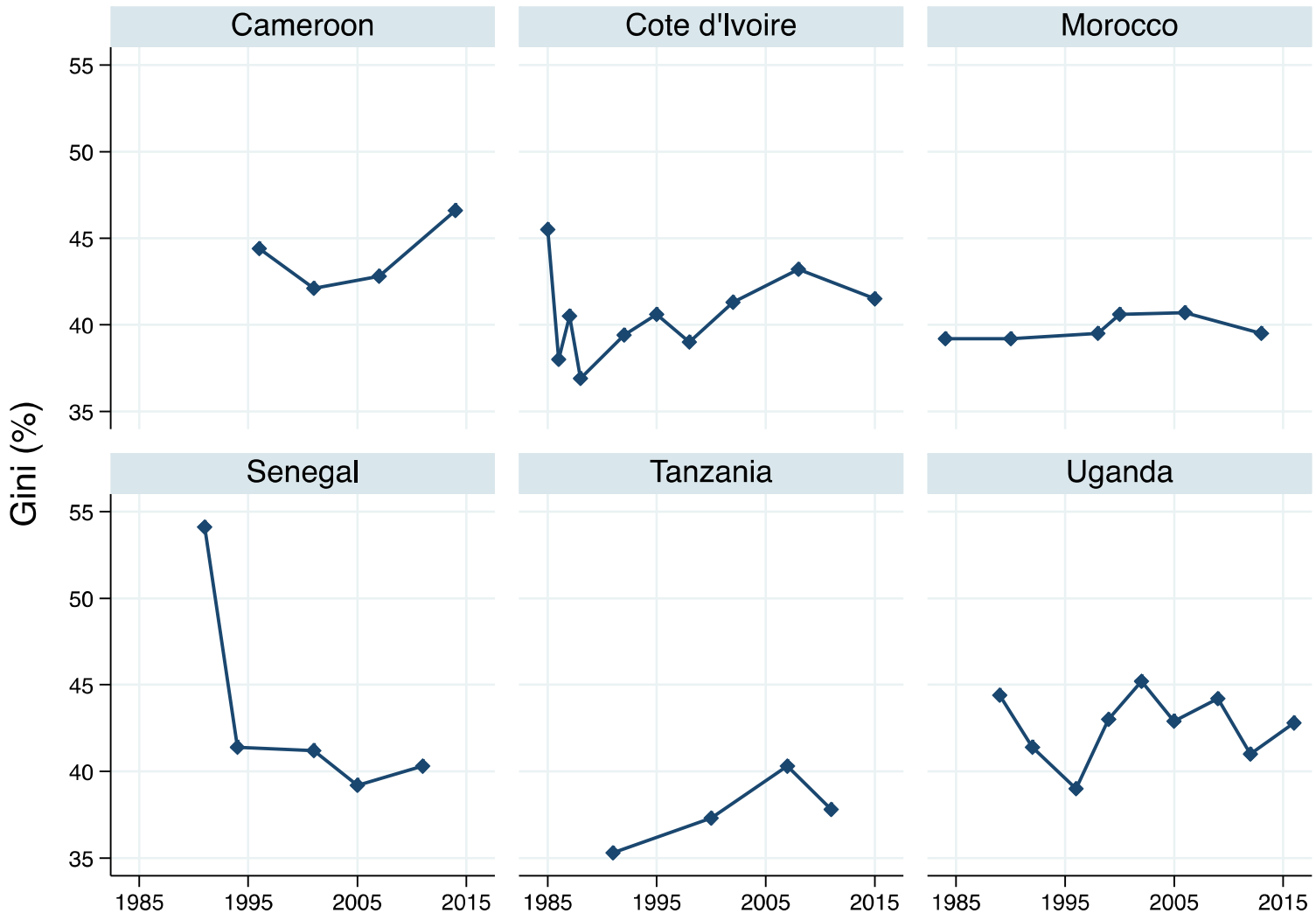
orange = expenditure, blue = income



Time trends

How robust is the inequality time trend for a given country?

Gini Index, WDI 1985-2015



What can go wrong?

Can you think of any factors that threaten the robustness of our findings?

1. Definition of ‘wealth’ can be different across countries
2. Data collection method can change over time
3. Data issues

1. Definitions

$$W = \sum_{j=1}^{k} \pi_j A_j - D$$

where:

W denotes wealth or 'net worth'

$A_j \geq 0$ is the amount of asset type j

π_j is the price of asset j

D is debt

Note: W can be negative

2. Data collection method

Journal of Development Economics 98 (2012) 3–18

Contents lists available at SciVerse ScienceDirect

Journal of Development Economics

journal homepage: www.elsevier.com/locate/devec

Methods of household consumption measurement through surveys: Experimental results from Tanzania ☆

Kathleen Beegle [a,*], Joachim De Weerdt [b], Jed Friedman [a], John Gibson [c]

[a] World Bank, United States
[b] EDI, Tanzania
[c] University of Waikato, New Zealand

ARTICLE INFO

Article history:
Received 30 January 2010
Received in revised form 1 November 2011
Accepted 2 November 2011

JEL classification:
C83
D12

Keywords:
Consumption
Expenditure
Survey design

ABSTRACT

Surveys of consumption expenditure vary widely across many dimensions, including the level of reporting, the length of the reference period, and the degree of commodity detail. These variations occur both across countries and also over time within countries, with little current understanding of the implications of such changes for spatially and temporally consistent measurement of household consumption and poverty. A field experiment in Tanzania tests eight alternative methods of measuring household consumption, finding significant differences between consumption reported by the benchmark personal diary and other diary and recall formats. Under-reporting is particularly apparent for illiterate households and for urban respondents completing household diaries; recall modules measure lower consumption than a personal diary, with larger gaps among poorer households and for households with more adult members. Variations in reporting accuracy by household characteristics are also discussed and differences in measured poverty as a result of survey design are explored.

Beegle et al. (2012) experiments

- "Our survey experiment entailed fielding eight alternative consumption questionnaires randomly assigned to 4,000 households in Tanzania."

- The eight designs vary by method of data capture, level of respondent, length of reference period, number of items in the recall list, and nature of the cognitive task required of the respondent.

Consumption expenditure per capita (annualized Tanzania shillings) by consumption module

Module	Mean Total
1. Long 14 day	476,721
2. Long 7 day	510,920
3. Subset 7 day	480,678
4. Collapse 7 day	401,925
5. Long usual 12 month	473,884
6. HH diary frequent	403,759
7. HH diary infrequent	416,043
8. Personal diary	500,702
All modules	467,840

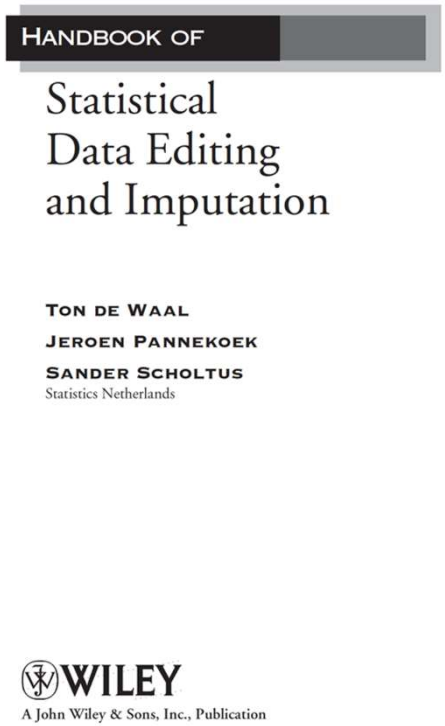
510,920 / 401,925 = 1.27

3. Data issues

- The process of data collection has inherent flaws.
- Data validation is a complex activity aimed at verifying that data intended for analytical purposes are cleaned and consistently organized into datasets

Data validation

1. Range checks
   simplest edits one can think of
2. Internal consistency checks
   combination of edits
3. Missing values
4. Outliers
   investigation of extreme values



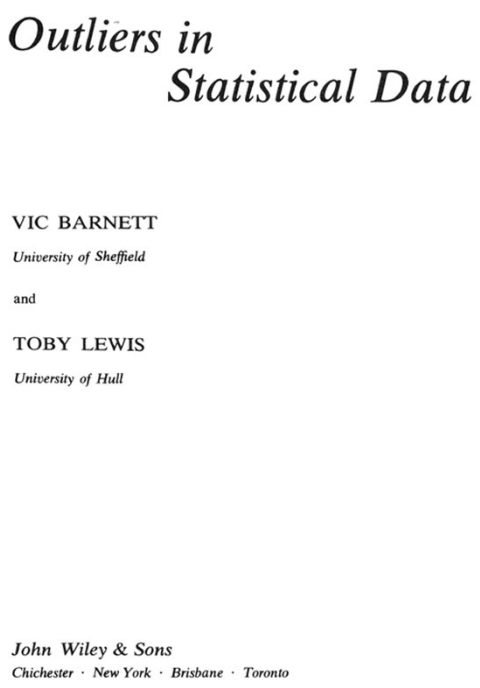
Outliers

Outlier?

A definition

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

- Barnett and Lewis (1978)



Outliers in Statistical Data

VIC BARNETT
University of Sheffield

and

TOBY LEWIS
University of Hull

John Wiley & Sons
Chichester · New York · Brisbane · Toronto

Outlier detection: does it matter?

- Theory first.
- Three papers:

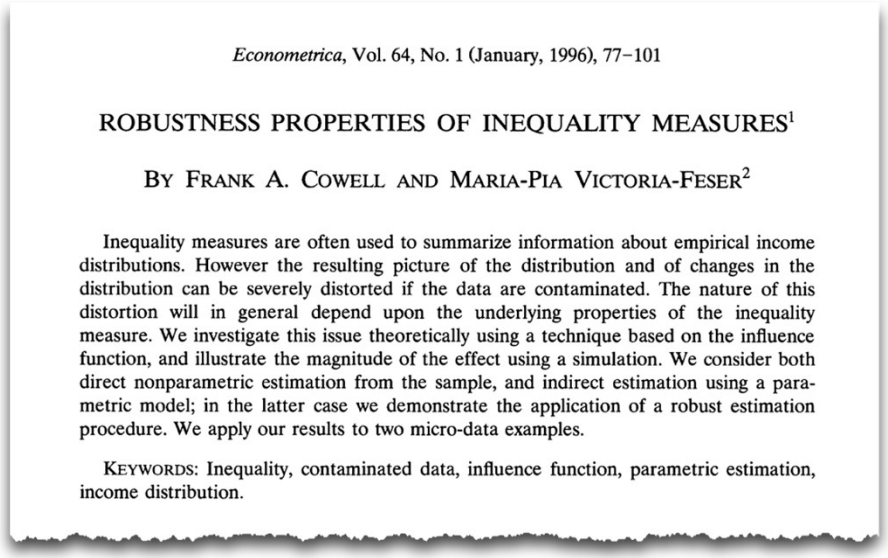
I. 1996a
   Frank Cowell and Maria-Pia Victoria-Feser

II. 2007
   Frank Cowell and Emmanuel Flachaire

III. 1996b
   Frank Cowell and Maria-Pia Victoria-Feser

Outliers and inequality measures – I

Cowell and Victoria-Feser (1996a)



Econometrica, Vol. 64, No. 1 (January, 1996), 77–101

ROBUSTNESS PROPERTIES OF INEQUALITY MEASURES[1]

BY FRANK A. COWELL AND MARIA-PIA VICTORIA-FESER[2]

Inequality measures are often used to summarize 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 nature of this distortion will in general depend upon the underlying properties of the inequality measure. We investigate this issue theoretically using a technique based on the influence function, and illustrate the magnitude of the effect using a simulation. We consider both direct nonparametric estimation from the sample, and indirect estimation 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 examples.

KEYWORDS: Inequality, contaminated data, influence function, parametric estimation, income 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.

The influence function

- F — Ideal data, no contaminants
- $I(F)$ — "true" Gini index

- $G = (1 - \delta)F + \delta H$ — Real-world data, with δ% contaminants
  $0 \leq \delta \leq 1$
- $I(G)$ — "estimated" Gini index

- The influence function, IF: $IF = \lim_{\delta \to 0} \frac{I(G) - I(F)}{\delta}$

The catastrophe

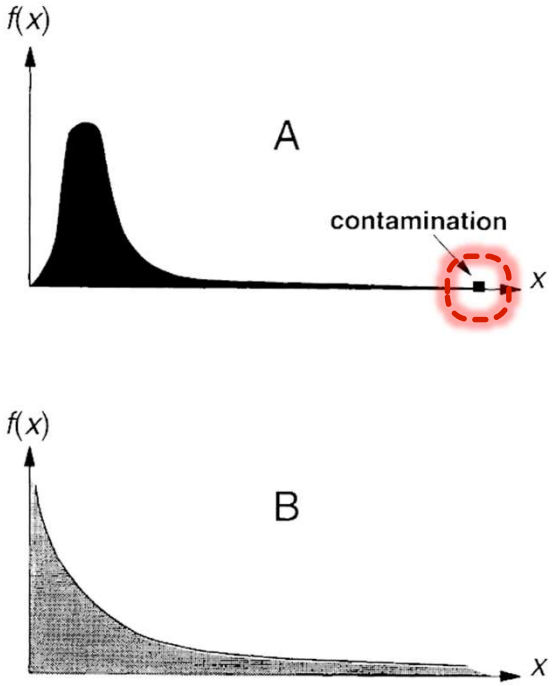


FIGURE 1

- Suppose the shape of the income distribution is represented by the continuous frequency distribution in part A
- Suppose that in the sample there are some rogue observations represented by the point mass labelled "contamination".
- Then, according to inequality statistics that are sensitive to the top end of the distribution, the income distribution in A will be indistinguishable from that represented in B (that is, IF is unbounded).

Do-it-yourself….

English

1) Generate a log-normal looking wealth distribution

2) Estimate the Gini index

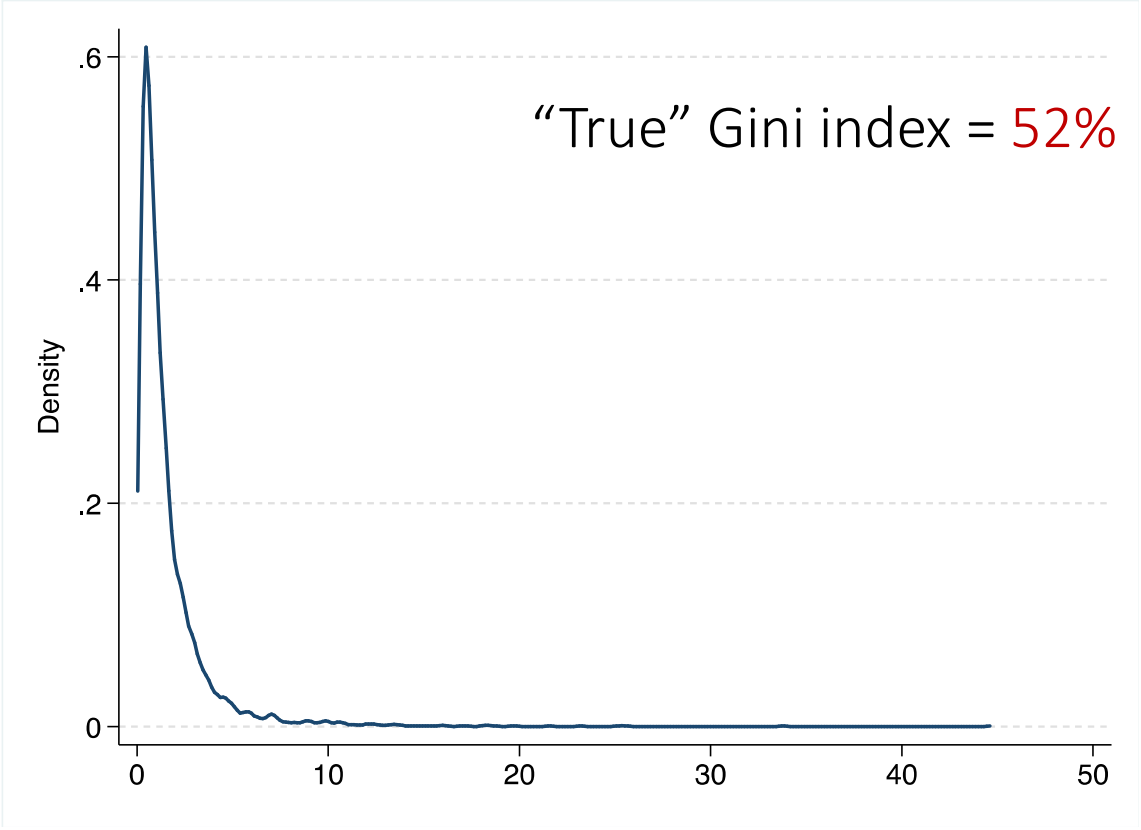
3) Contaminate the distribution with a few extreme values

4) Re-estimate the Gini index

Stata/R/SPSS/Excel/…

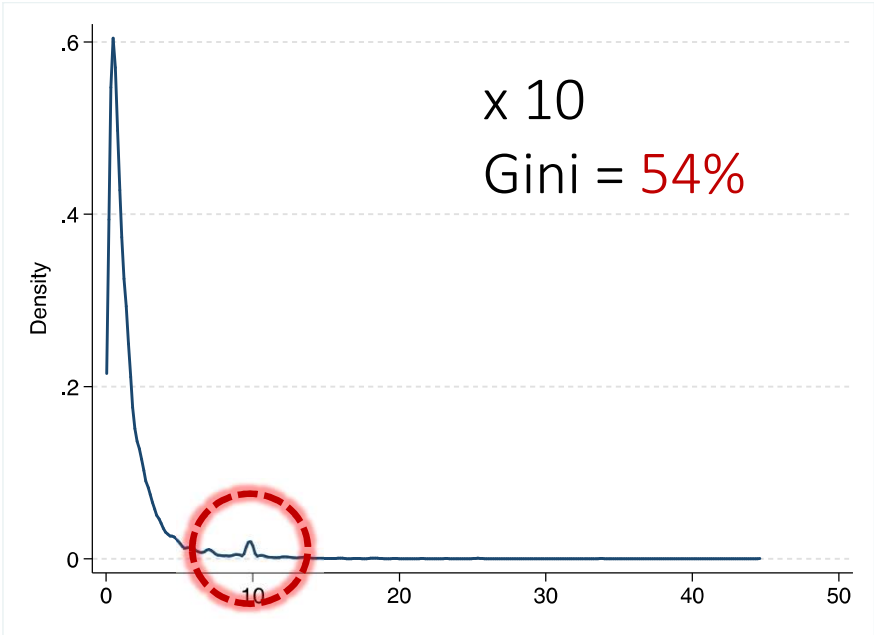
```
clear

set obs 5000
set seed 198607
gen n  = rnormal(0,1)
gen ln = exp(n)

* simulate order of magnitude mistake:
* take 100 obs around the median
* of the distribution and multiply
* them by 100

sort ln

gen cont100 = 1
replace cont100 = 100 in 2480/2520
gen ln_cont100 = ln*cont100
```

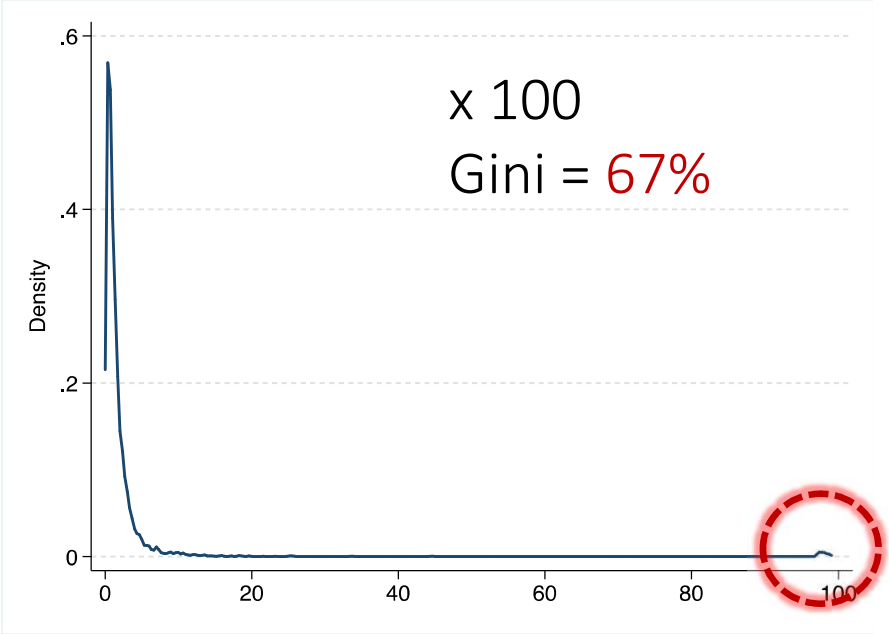
“True” wealth distribution



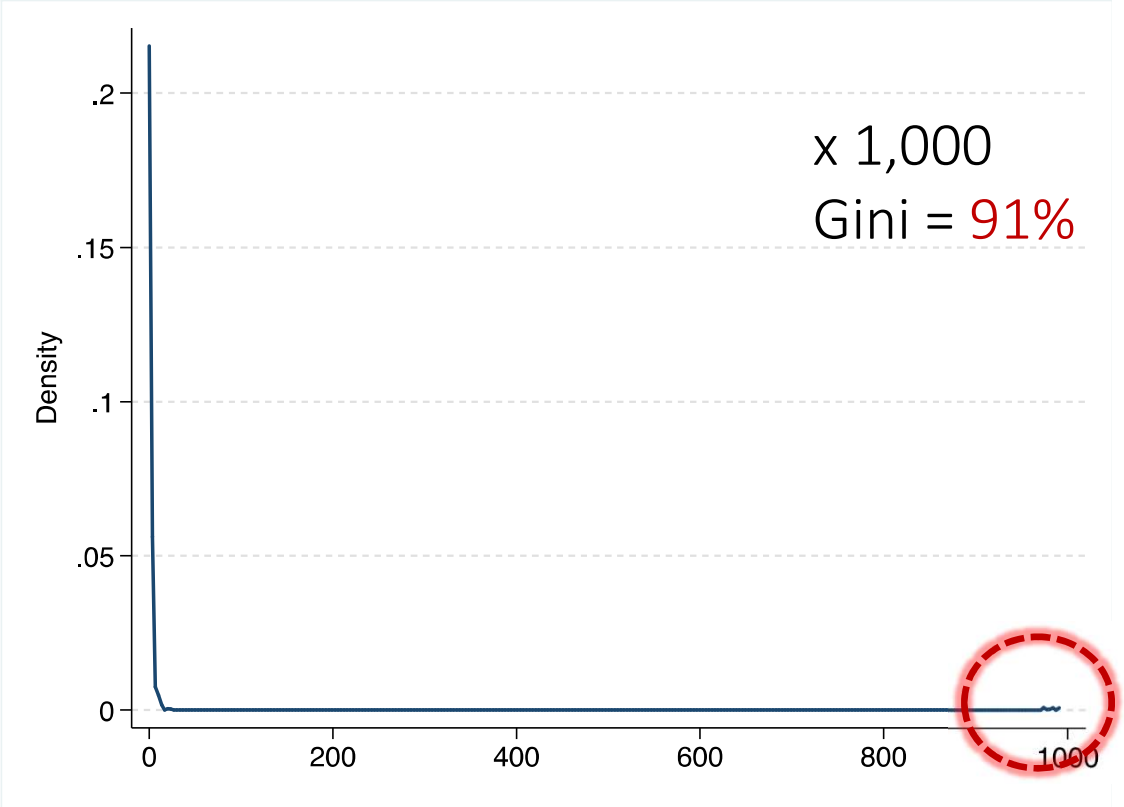
Contamination

40 out of 5,000 observations (less than 1%) are ”contaminated”



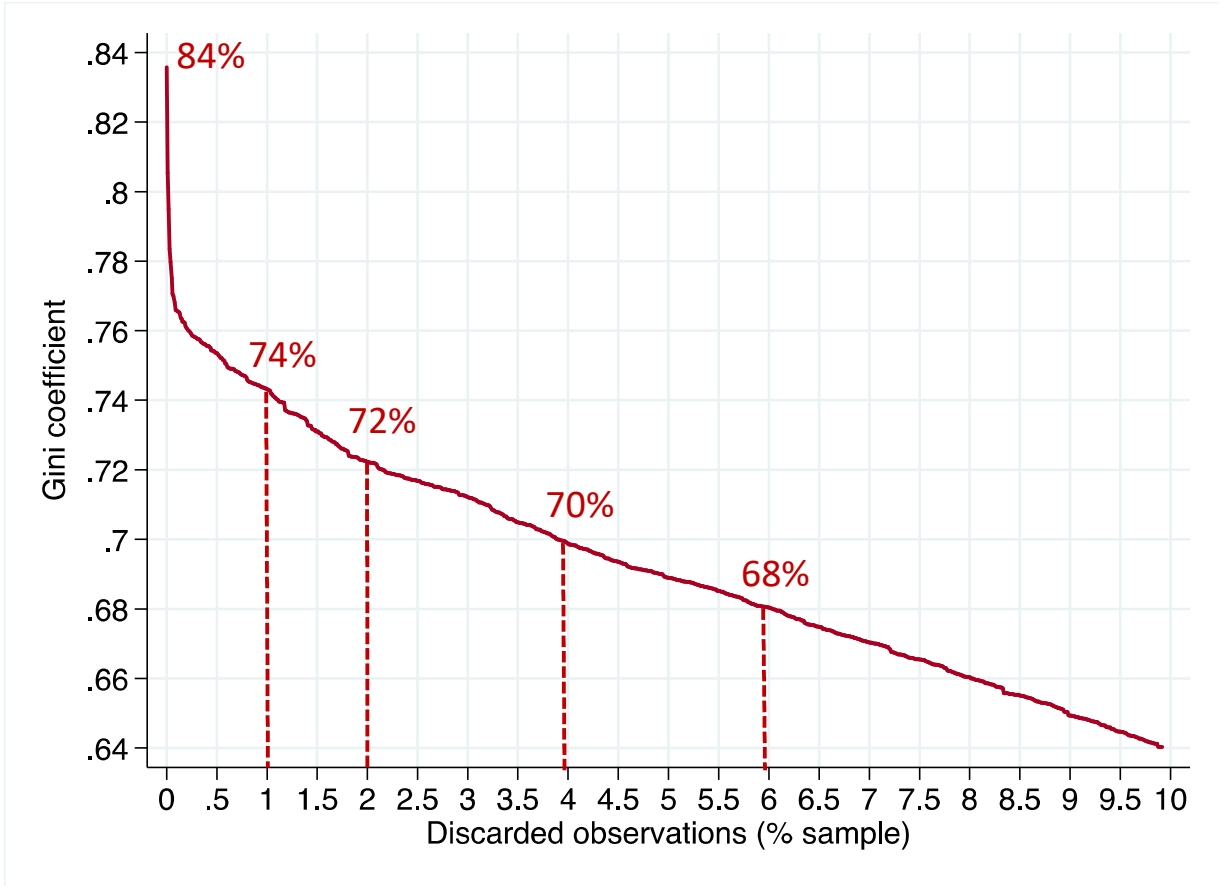


Contamination



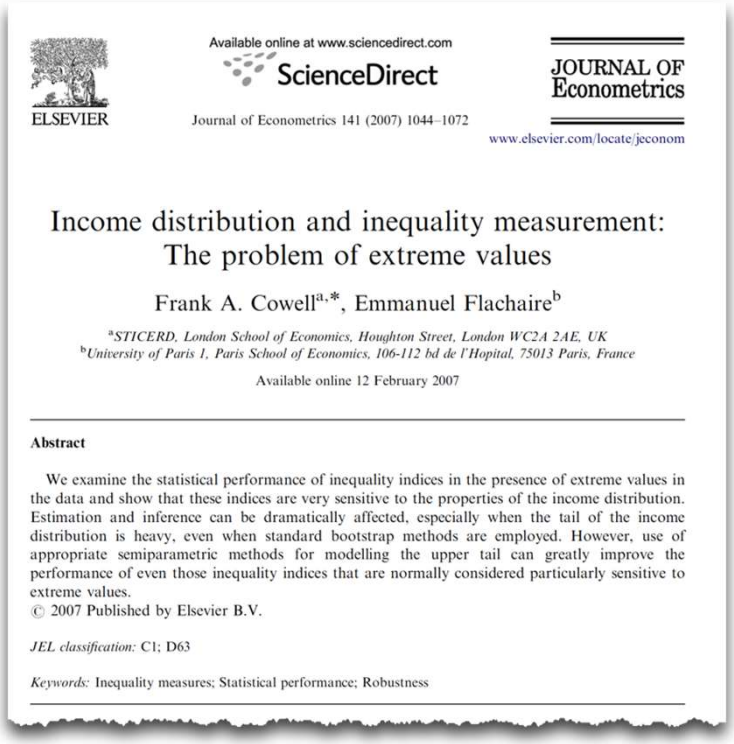
Sensitivity of the Gini index to extreme values

cumulative truncation



Outliers and inequality measures – II

Cowell and Flachaire (2007)



Available online at www.sciencedirect.com
ScienceDirect
Journal of Econometrics 141 (2007) 1044–1072
JOURNAL OF Econometrics
www.elsevier.com/locate/jeconom

Income distribution and inequality measurement: The problem of extreme values

Frank A. Cowell[a,*], Emmanuel Flachaire[b]

[a]STICERD, London School of Economics, Houghton Street, London WC2A 2AE, UK
[b]University of Paris 1, Paris School of Economics, 106-112 bd de l'Hopital, 75013 Paris, France

Available online 12 February 2007

Abstract

We examine the statistical performance of inequality indices in the presence of extreme values in the data and show that these indices are very sensitive to the properties of the income distribution. Estimation and inference can be dramatically affected, especially when the tail of the income distribution is heavy, even when standard bootstrap methods are employed. However, use of appropriate semiparametric methods for modelling the upper tail can greatly improve the performance of even those inequality indices that are normally considered particularly sensitive to extreme values.

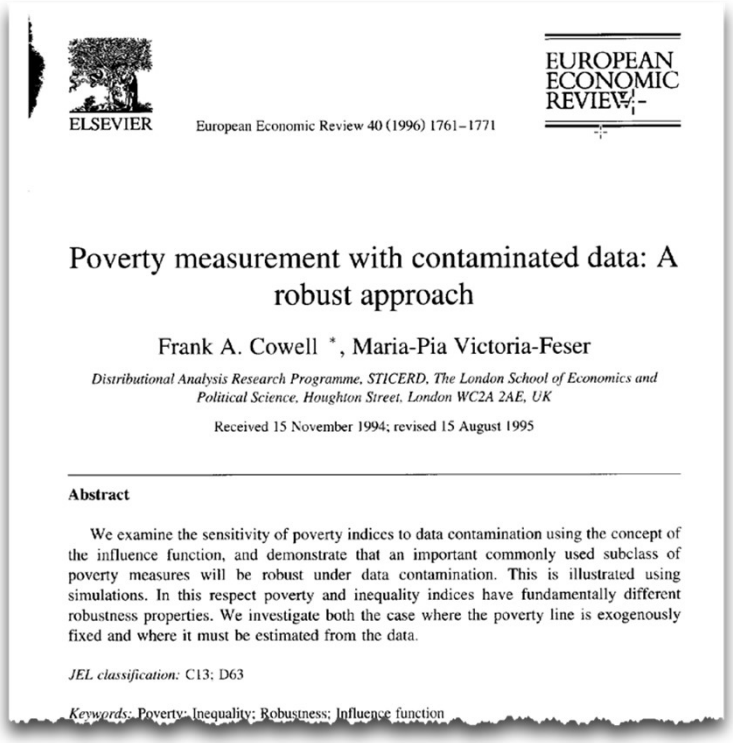
JEL classification: C1; D63

Keywords: Inequality measures; Statistical performance; Robustness

- Explains how and why outliers are a serious threat to most inequality measures.
- Suggests to use the ECDF for all but the right-hand tail + parametric estimation for the upper tail

Outliers and poverty measures

Cowell and Victoria-Feser (1996b)



ELSEVIER

European Economic Review 40 (1996) 1761–1771

EUROPEAN ECONOMIC REVIEW

Poverty measurement with contaminated data: A robust approach

Frank A. Cowell *, Maria-Pia Victoria-Feser

Distributional Analysis Research Programme, STICERD, The London School of Economics and Political Science, Houghton Street, London WC2A 2AE, UK

Received 15 November 1994; revised 15 August 1995

Abstract

We examine the sensitivity of poverty indices to data contamination using the concept of the influence function, and demonstrate that an important commonly used subclass of poverty measures will be robust under data contamination. This is illustrated using simulations. In this respect poverty and inequality indices have fundamentally different robustness properties. We investigate both the case where the poverty line is exogenously fixed and where it must be estimated from the data.

JEL classification: C13; D63

Keywords: Poverty; Inequality; Robustness; Influence function

- Explains why outliers only rarely are a serious threat to most poverty measures.
- In a nutshell, if the poverty line is exogenous, the poverty measures are not sensitive to the values (real or contaminated) of the incomes of the rich

Recap

- Edits
  documentation and replicability, otherwise comparisons are going to be inconsistent
- Outliers
  both theory (unbounded IF) and practice (cumulated truncation) suggest that they matter (tremendously)

Outlier detection

- The literature is rich with methods to identify outliers; in practice, most methods used in empirical works hinge on the underlying distribution of the data.

- The idea is simple:
  - Transform the variable to induce normality
  - Set thresholds to identify extreme values

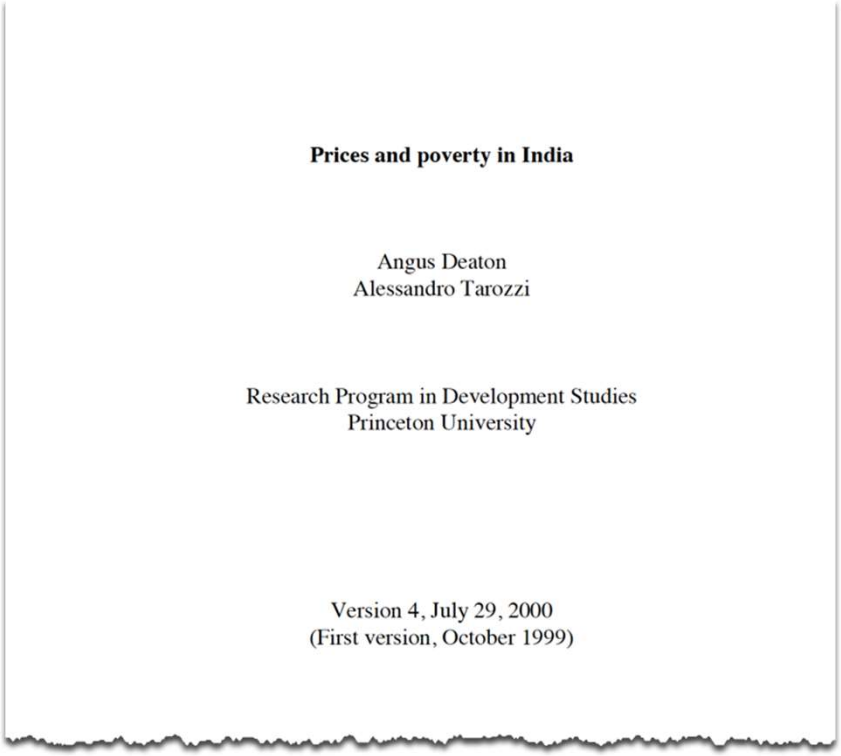
Transform the variable to induce normality

- A classical transformation relies on z-scores:

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

where $\bar{x}$ is the mean and s is the standard deviation

Deaton and Tarozzi (2000)



Prices and poverty in India

Angus Deaton
Alessandro Tarozzi

Research Program in Development Studies
Princeton University

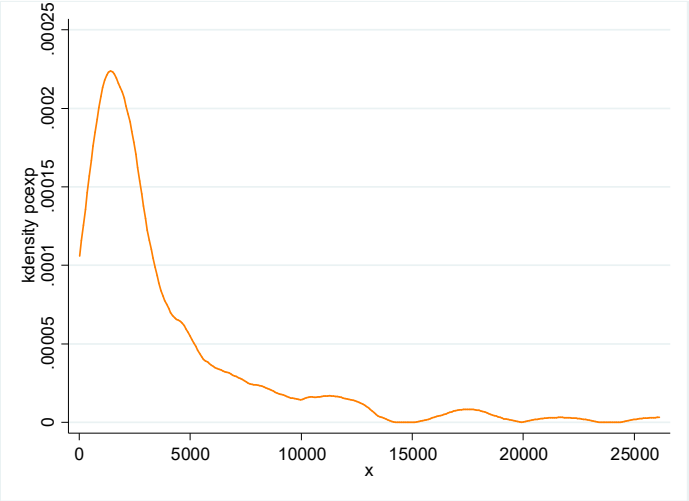
Version 4, July 29, 2000
(First version, October 1999)

In the case of India, D&T (2000) flagged as outliers prices whose logarithms exceeded the mean of logarithms by more than 2.5 standard deviations:

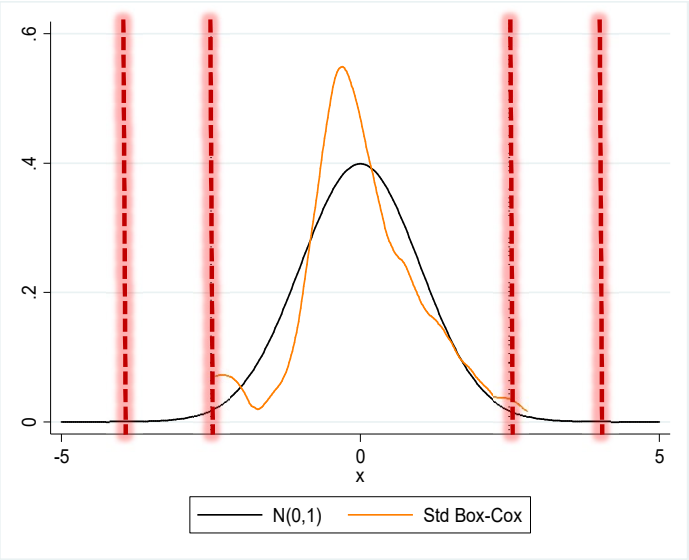
$$\frac{ln(x) - E[ln(x)]}{sd[ln(x)]} > 2.5$$

Transformation and normalization

Raw untransformed data



Transformed data



Two questions

1) How good is such an approach?

2) What to do after flagging outliers?

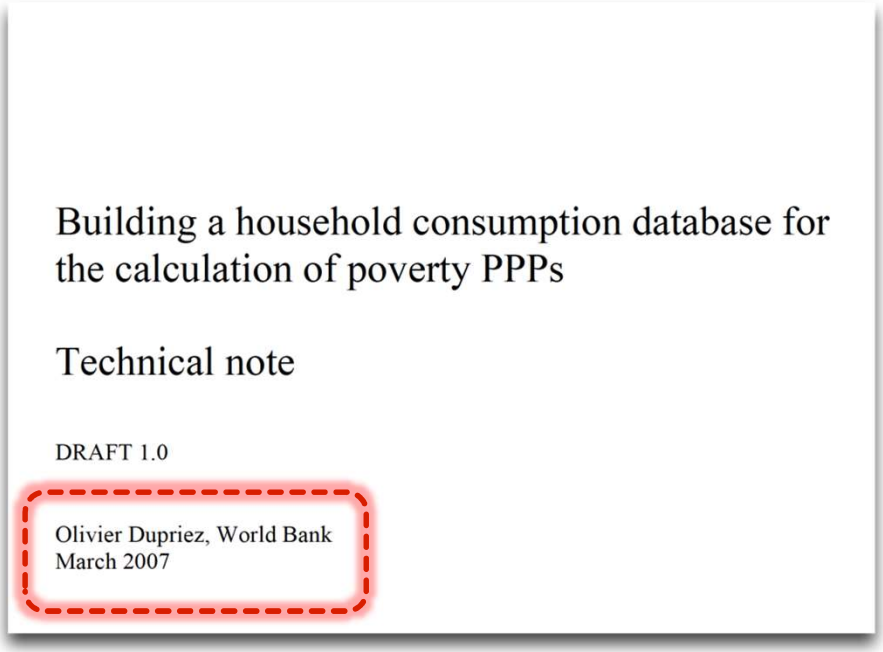
How good is such an approach?

- Log-transformation is very basic – how to deal with negative values?
- Why using mean and standard deviation?

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

- Not robust
- We can do better

The Box-Cox transformation



Building a household consumption database for the calculation of poverty PPPs

Technical note

DRAFT 1.0

Olivier Dupriez, World Bank
March 2007

- The Box-Cox transformation:

$$y_h^{(\lambda)} = \begin{cases} (y_h^{\lambda} - 1)/\lambda & \text{if } \lambda \neq 0 \\ \ln y_h & \text{if } \lambda = 0 \end{cases}$$

- Outliers are identified if:

y_h > 75th percentile + 5 × IQR

The median absolute deviation (MAD)

$$z_h = \frac{x_h - \bar{x}}{s} \qquad z_h = \left| \frac{x_h - med[x_h]}{MAD} \right|$$

$$MAD = b \times med|x - med[x]|$$

b = 1.4826

if the distribution is Gaussian

We can do 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 $\mathrm{MAD}_n = 1.4826\ \mathrm{med}_i\{|x_i - \mathrm{med}_j x_j|\}$, 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_n 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 $S_n = 1.1926\ \mathrm{med}_i\{\mathrm{med}_j|x_i - x_j|\}$ and the estimator Q_n given by the .25 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 (for implosion as well as explosion), 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.

Rousseeuw and Croux (1993)

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

- $S = c \times \mathrm{med}_i\{\mathrm{med}_j|x_j - x_i|\}$

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

$$z_h = \left|\frac{x_h - med[x_h]}{S}\right|$$

c = 1.1926 at the Gaussian model.

Treatment of outliers

Three main methods of 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)
3) using robust estimation techniques (M-estimation).

- Documentation, transparency & reproducibility

One last example

OECD (2013)

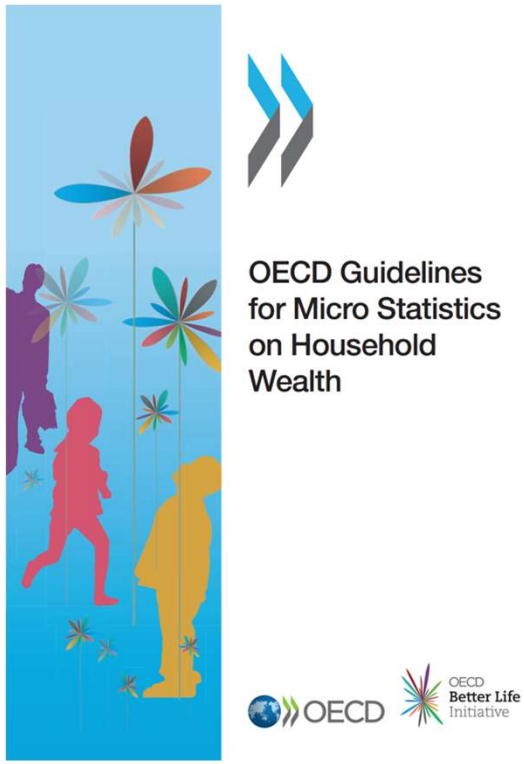


Table 7.3. **Effect of the treatment of outliers on summary measures of wealth inequality in the United States, 2007**

	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
n	4 418	3 698	4 359

Source: 2007 Survey of Consumer Finances.

Recap

- Detection
  - “take the log and run” is not a recommended practice
  - MAD (median absolute deviation)
  - Rousseeuw and Croux (1993)
- Treatment
  - no consensus
  - quantile regression?

Conclusions

1) Editing rules – take it seriously and document them
   replicability

2) As far as inequality is concerned, outliers are the worst enemy
   Cowell and Victoria-Feser (1996): unbounded IF

3) Outlier detection and treatment
   beyond logs and Box-Cox transformations
   Rousseeuw and Croux (1993): robustified scores

All this having been said

- Outliers can be genuine observations…
- Be gentle to the data and document each and every step of the data processing

Thank you for your attention

References

- Barnett and Lewis (1974). *Outliers in statistical data*, John Wiley & Sons.
- Beegle, De Weerdt, Friedman, and Gibson (2012). Methods of household consumption measurement through surveys: Experimental results from Tanzania. *Journal of Development Economics, Elsevier,* vol. 98(1), pages 3-18
- Cowell and Flachaire (2007). Income distribution and inequality measurement: The problem of extreme values. *Journal of Econometrics*, *141*(2), 1044-1072.
- Cowell and Victoria-Feser (1996a). Robustness properties of inequality measures. *Econometrica: Journal of the Econometric Society*, 77-101.
- Cowell and Victoria-Feser (1996b). Poverty measurement with contaminated data: A robust approach. *European Economic Review*, *40*(9), 1761-1771.
- De Waal, Pannekoek, and Scholtus (2011). *Handbook of statistical data editing and imputation*, John Wiley & Sons.
- Deaton and Tarozzi (2005). Prices and Poverty in India. *The Great Indian Poverty Debate*. New Delhi : MacMillan.
- Organisation for Economic Co-operation and Development (2013). *OECD Guidelines for Micro Statistics on Household Wealth*. OECD Publishing.
- Rousseeuw and Croux (1993). Alternatives to the median absolute deviation. *Journal of the American Statistical association*, *88*(424), 1273-1283.