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


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)


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

Cameroon

Cote d'Ivoire

Morocco

Senegal

Tanzania

Uganda
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 \( j \)
- \( \pi_j \) is the price of asset \( j \)
- \( D \) is debt

Note: \( W \) can be negative
2. Data collection method
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

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

\[
\frac{510,920}{401,925} = 1.27
\]
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)
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)

- 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

- $F$
- $I(F)$
- $G = (1 - \delta)F + \delta H$
  \[0 \leq \delta \leq 1\]
- $I(G)$

The influence function, $IF$:

$$IF = \lim_{{\delta \to 0}} \frac{I(G) - I(F)}{\delta}$$

Ideal data, no contaminants

“true” Gini index

Real-world data, with $\delta\%$ contaminants

“estimated” Gini index
The catastrophe

- 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....

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

```plaintext
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

“True” Gini index = 52%
Contamination

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

\[ \times 10 \quad \text{Gini} = 54\% \]

\[ \times 100 \quad \text{Gini} = 67\% \]
Contamination

Gini = 91%

x 1,000
Sensitivity of the Gini index to extreme values
cumulative truncation
Outliers and inequality measures – II
Cowell and Flachaire (2007)

- 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)

- 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
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

![Graph showing raw untransformed data and transformed data.](image-url)
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

- 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:

\[y_h > 75\text{th percentile} + 5 \times \text{IQR}\]
The median absolute deviation (MAD)

\[ z_h = \frac{x_h - \bar{x}}{s} \quad \text{or} \quad z_h = \left| \frac{x_h - med[x_h]}{MAD} \right| \]

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

\[ b = 1.4826 \]

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 \( \text{MAD}_n = 1.4826 \text{med}_i \{ |x_i - \text{med}_j x_j| \} \), 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 \( \text{MAD}_n \) 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 \( S_n = 1.1926 \text{med}_i \{ \text{med}_j |x_i - x_j| \} \) 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) propose to substitute the MAD with a different estimator:

\[ S = c \times \text{med}_i \{ \text{med}_j |x_j - x_i| \} \]

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 \( S \).

\[ z_h = \left| x_h - \text{med}[x_h] \right| / S \quad \Rightarrow \quad c = 1.1926 \text{ at the Gaussian model.} \]
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**

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

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