Outlier detection and treatment

LECTURE 12
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?
Definitions
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)

- Note: we focus on univariate outliers, those found when looking at a distribution of values in a single dimension (e.g. income).
Highest sea-levels in Venice
Other classical definitions

- An outlier is “an observation that deviates so much from other observations as to arouse suspicion that it was generated by a different mechanism” (Hawkins 1980)

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
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
Do outliers 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 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).
In practice
Hlasny and Verme (2018: 191)

- Many researchers routinely \textit{trim} outliers or problematic observations or apply \textit{top coding} with little consideration of the implications for the measurement of inequality

- One example to illustrate
Sensitivity of the Gini index to extreme values

iterative trimming
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.
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
How to detect outliers?
Visual inspection

- 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 (Deaton and Tarozzi 2005)
Visual inspection
Malawi IHS3, Cassava tuber expenditure
Visual inspection
Malawi IHS3, Cassava tuber expenditure

- Example 1: look at descriptive statistics

<table>
<thead>
<tr>
<th>Percentiles</th>
<th>Smallest</th>
</tr>
</thead>
<tbody>
<tr>
<td>1%</td>
<td>5</td>
</tr>
<tr>
<td>5%</td>
<td>20</td>
</tr>
<tr>
<td>10%</td>
<td>20</td>
</tr>
<tr>
<td>25%</td>
<td>50</td>
</tr>
<tr>
<td>50%</td>
<td>75</td>
</tr>
<tr>
<td>75%</td>
<td>100</td>
</tr>
<tr>
<td>90%</td>
<td>200</td>
</tr>
<tr>
<td>95%</td>
<td>220</td>
</tr>
<tr>
<td>99%</td>
<td>350</td>
</tr>
</tbody>
</table>

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

How much did you spend?

<table>
<thead>
<tr>
<th></th>
<th>Smallest</th>
</tr>
</thead>
<tbody>
<tr>
<td>Obs</td>
<td>673</td>
</tr>
<tr>
<td>Sum of Wgt.</td>
<td>673</td>
</tr>
<tr>
<td>Largest</td>
<td></td>
</tr>
<tr>
<td>Mean</td>
<td>94.95097</td>
</tr>
<tr>
<td>Std. Dev.</td>
<td>106.2979</td>
</tr>
<tr>
<td>Variance</td>
<td>11286.5</td>
</tr>
<tr>
<td>Skewness</td>
<td>10.0151</td>
</tr>
<tr>
<td>Kurtosis</td>
<td>164.7054</td>
</tr>
</tbody>
</table>

C4D2 Training
Visual inspection
Malawi IHS3, Cassava tuber expenditure

- Example 2: graph the distribution of the data
Visual inspection
Malawi IHS3, Cassava tuber expenditure

- Example 3: use graphical diagnostic tools, e.g. the boxplot graph
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
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.
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)
Why 2.5, 3, 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 other observations normally are.
Deaton and Tarozzi (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:

\[
\frac{\ln(x) - E[\ln(x)]}{sd[\ln(x)]} > 2.5
\]
Transformation and thresholds

Raw untransformed data

Transformed data

\[ \text{kdensity pcexp} \]

\[ x \]

\[ N(0,1) \] Std Box-Cox
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?
- Not recommended when the log-distribution can not be assumed to be a Normal distribution
- 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
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).
The median absolute deviation (MAD)

\[ z_h = \frac{x_h - \bar{x}}{s} \]

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

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

\[ b = 1.4826 \]

if the distribution is Gaussian
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 \( \text{MAD}_n = 1.4826 \ \text{med} \{ |x_i - \text{med} x_j| \} \), 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 \( S_n = 1.1926 \ \text{med} \{ \text{med}_i |x_i - x_j| \} \) 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 \( |x_i - x_j| \) (\( j = 1, \ldots, n \)). This yields \( n \) numbers, the median of which gives our final estimate \( S \).

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

\( c = 1.1926 \) at the Gaussian model.
Recap

- “take the log and run” is not a recommended practice
- taking the log and robustifying the z-score is a better practice
- Belotti and Vecchi (2019) provide `outdetect.ado`
Malawi, 2013

- ‘take the log and run’: 2.08% of outliers (most of which in the right tail)
- ‘take the log, robustify the z-score, and run’: 3.00% (most of which in the right tail)

<table>
<thead>
<tr>
<th>Method</th>
<th>Overall</th>
<th>Left</th>
<th>Right</th>
</tr>
</thead>
<tbody>
<tr>
<td>Z-score</td>
<td>2.08</td>
<td>0.20</td>
<td>1.88</td>
</tr>
<tr>
<td>MAD-score</td>
<td>3.05</td>
<td>0.35</td>
<td>2.70</td>
</tr>
<tr>
<td>S-score</td>
<td>3.02</td>
<td>0.35</td>
<td>2.67</td>
</tr>
<tr>
<td>Q-score</td>
<td>3.00</td>
<td>0.35</td>
<td>2.65</td>
</tr>
</tbody>
</table>
How to deal with outliers?
(in one slide)
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
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.
  - Use a “*take the log, robustify the z-score and run*”, strategy.

- **Outlier treatment**: it depends. Quantile regression is a good candidate.
References

Required readings

Suggested readings
Thank you for your attention
Homework
Exercise 1 - Engaging with the literature

Summarize the main conclusions of the paper: do outliers matter? Why or why not?
Exercise 2 - 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

---

**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
```
Exercise 3 – Inequality measures

- Comment on table 7.3 from OECD (2013) p.172 (see next slide).
- What can you say about the sensitivity of estimates to the treatment of outliers?
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

<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>Mean</td>
<td>556 846</td>
<td>378 215</td>
<td>559 361</td>
</tr>
<tr>
<td>Median</td>
<td>120 780</td>
<td>120 780</td>
<td>123 800</td>
</tr>
<tr>
<td>Gini</td>
<td>0.82</td>
<td>0.74</td>
<td>0.81</td>
</tr>
<tr>
<td>$% CV^2$</td>
<td>18.1</td>
<td>2.4</td>
<td>14.6</td>
</tr>
<tr>
<td>P90/P10</td>
<td>30 000</td>
<td>3 369</td>
<td>3 061</td>
</tr>
<tr>
<td>P75/P25</td>
<td>26.3</td>
<td>24.5</td>
<td>24.3</td>
</tr>
<tr>
<td>P90/P50</td>
<td>7.6</td>
<td>7.0</td>
<td>7.4</td>
</tr>
<tr>
<td>$n$</td>
<td>4 418</td>
<td>3 698</td>
<td>4 359</td>
</tr>
</tbody>
</table>