Outlier detection and treatment

LECTURE 12

C4D2 TRAINING

1

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?

C4D2 TRAINING

2

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

C4D2 TRAINING

3

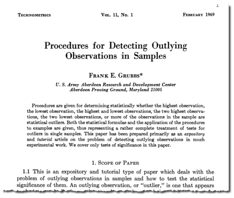
3

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)


TECHNOMETRICS VOL. 11, NO. 1 FEBRUARY 1969

Procedures for Detecting Outlying Observations in Samples

FRANK E. GRUBBS*

U. S. Army Aberdeen Research and Development Center
Aberdeen Proving Ground, Maryland 21005

Procedures are given for determining statistically whether the highest observation, the lowest observation, the highest and lowest observations, the two highest observations, the two lowest observations, or more of the observations in the sample are statistical outliers. Both the statistical formulae and the application of the procedures to examples are given, thus representing a rather complete treatment of tests for outliers in single samples. This paper has been prepared primarily as an *expository* and *tutorial* article on the problem of detecting outlying observations in much experimental work. We cover only tests of significance in this paper.

1. SCOPE OF PAPER

1.1 This is an expository and tutorial type of paper which deals with the problem of outlying observations in samples and how to test the statistical significance of them. An outlying observation, or “outlier,” is one that appears

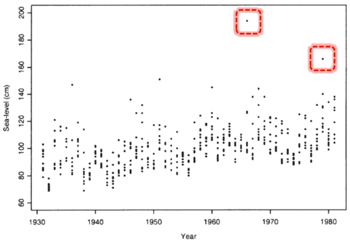
5

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

Highest sea-levels in Venice



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

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

9

Do outliers matter?

C4D2 TRAINING 10

10

Theory first

- Three papers:

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

  II. 2007
  Frank Cowell and Emmanuel Flachaire (*)

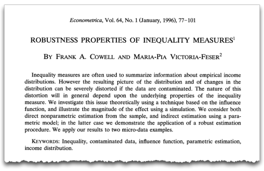
  III. 1996b
  Frank Cowell and Maria-Pia Victoria-Feser

C4D2 TRAINING

11

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.

C4D2 TRAINING

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.

C4D2 TRAINING 15

15

In practice

Hlasny and Verme (2018: 191)

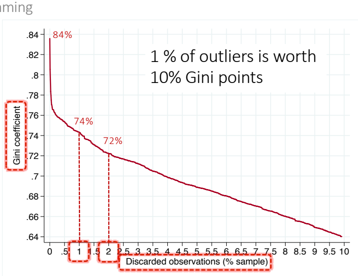
- Many researchers routinely trim outliers or problematic observations or apply top coding with little consideration of the implications for the measurement of inequality
- One example to illustrate

C4D2 TRAINING 16

16

Sensitivity of the Gini index to extreme values

iterative trimming

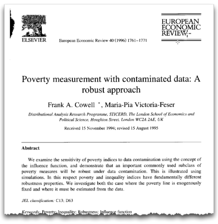


C4D2 TRAINING

17

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.
- Poverty measures are not sensitive to the values (real or contaminated) of the incomes of the rich

C4D2 TRAINING

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

C4D2 TRAINING

21

How to detect outliers?

C4D2 TRAINING 22

22

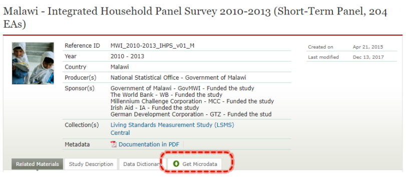
Visual inspection
(Deaton and Tarozzi 2005)

- "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"

C4D2 TRAINING

23

Visual inspection
Malawi IHS3, Cassava tuber expenditure



C4D2 TRAINING

24

Visual inspection
Malawi IHS3, Cassava tuber expenditure

MODULE G: FOOD CONSUMPTION OVER PAST ONE WEEK

DATA ENTRY LINE NUMBER	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.	G01 YES..1 NO...2>> NEXT ITEM	G02 ITEM CODE	G05 How much did you spend? MK
19	**Roots, Tubers, and Plantains**			
20	Cassava tubers		201	

C4D2 TRAINING

25

Visual inspection

Example 1: look at descriptive statistics

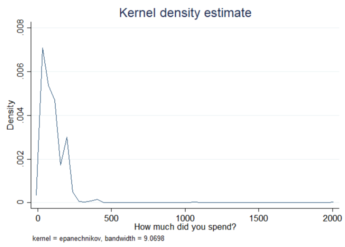
```
. sum hh_g05 if hh_g02==201,d

                    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
                        Largest       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
```

C4D2 TRAINING 26

26

Visual inspection

Example 2: graph the distribution of the data

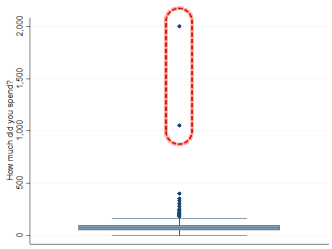


C4D2 TRAINING 27

27

Visual inspection

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



C4D2 TRAINING 28

28

Statistical methods

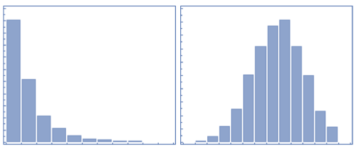
- The literature is rich with methods to identify outliers; in practice, most methods used in empirical work hinge on the underlying distribution of the data.
- The idea is simple:
  - transform the variable to induce normality
  - set thresholds to identify extreme values

C4D2 TRAINING

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.



C4D2 TRAINING

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 $x > x + z_\alpha \, s$

  where z_α equals, say, 2.5.
- We can express the same criterion as $\frac{x - x}{s} > z_\alpha$

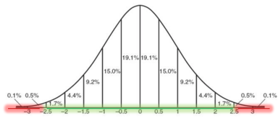
  where the left-hand side is called a z-score (a variable with mean = 0 and var = 1)

C4D2 TRAINING 31

31

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

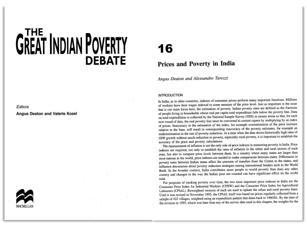
- Under the assumption of normality:



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

32

Deaton and Tarozzi (2005)



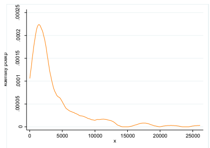
In the case of India, D&T (2005) flagged as outliers prices whose logarithms exceeded the mean of logarithms by more than 2.5 standard deviations:

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

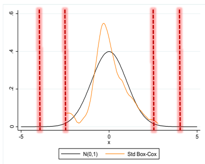
33

Transformation and thresholds

Raw untransformed data



Transformed data



34

Two questions

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

C4D2 Training

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

C4D2 Training

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 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).

C4D2 Training 39

39

Robustification of the z-score

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

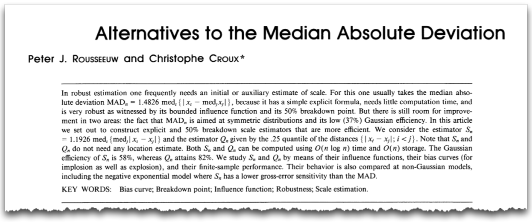
if the distribution is Gaussian

C4D2 Training

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 $MAD_n = 1.4826\, med_i \{|x_i - med_j x_j|\}$, 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 $S_n = 1.1926\, med_i \{med_j |x_i - x_j|\}$ 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.

C4D2 Training

41

Rousseeuw and Croux (1993)

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

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

- 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.

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

c = 1.1926 at the Gaussian model.

C4D2 Training

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` helps with both things

C4D2 TRAINING

43

Take the log and run vs. robust z-scores

Countries	Year	Outliers (%)					
cutoff = 3		log-transformation			robust z-scores		
		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

C4D2 TRAINING 44

44

Impact of outliers on the Gini index

Countries	Year	Gini index		
cutoff = 3				
		Raw	Trimmed (log)	Trimmed (best)
		(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

C4D2 TRAINING 45

45

How to deal with outliers?

(in one slide)

C4D2 TRAINING 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

C4D2 TRAINING

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.

C4D2 TRAINING 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., Garcıa-Fernández, R. M., Blanco-Encomienda, F. J., & Munoz, 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.

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

Cowell, F. A., & Flachaire, E. (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.

Deaton, A., & Tarozzi, 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/4YG7I5RGT0.

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

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

C4D2 TRAINING

49

Thank you for your attention

C4D2 TRAINING

50

50

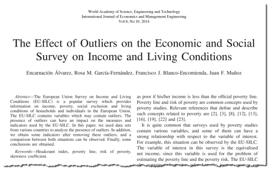
Homework

C4D2 TRAINING

51

51

Exercise 1 - Engaging with the literature



World Academy of Science, Engineering and Technology
International Journal of Economics and Management Engineering
Vol:8, No:10, 2014

The Effect of Outliers on the Economic and Social Survey on Income and Living Conditions

Encarnación Álvarez, Rosa M. García-Fernández, Francisco J. Blanco-Encomienda, Juan F. Muñoz

Abstract—The European Union Survey on Income and Living Conditions (EU-SILC) is a popular survey which provides information on income, poverty, social exclusion and living conditions of households and individuals in the European Union. The EU-SILC contains variables which may contain outliers. The presence of outliers can have an impact on the measures and indicators used by the EU-SILC. In this paper, we used data sets from various countries to analyze the presence of outliers. In addition, we obtain some indicators after removing these outliers, and a comparison between both situations can be observed. Finally, some conclusions are obtained.

Keywords—Headcount index, poverty line, risk of poverty, skewness coefficient.

as poor if his/her income is less than the official poverty line. Poverty line and risk of poverty are common concepts used by poverty studies. Relevant references that define and describe such concepts related to poverty are [2], [3], [8], [12], [13], [16], [19], [22] and [23].

It is quite common that surveys used by poverty studies contain various variables, and some of them can have a strong relationship with respect to the variable of interest. For example, this situation can be observed by the EU-SILC. The variable of interest in this survey is the equivalised net income, since this variable is used for the problem of estimating the poverty line and the poverty risk. The EU-SILC

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

C4D2 TRAINING

52

52

Exercise 2 - 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/...

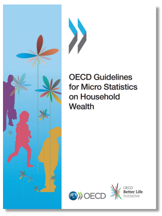
```
clear

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
```

C4D2 TRAINING

53

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?



C4D2 TRAINING

57

57

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
n	4 418	3 698	4 359

Source: 2007 Survey of Consumer Finances.

C4D2 TRAINING

58

58