UNIVERSITÄT
DUISBURG
ESSEN

Open-Minded

Outliers and Data Envelopment Analysis

Chair of Statistics
Prof. Dr. Andreas Behr
24.09.2024

Table of Contents

Conflicting interests in outlier detection

- Deleting the most efficient unit because of identification as outlier has important effects

1. The new benchmark has a lower productivity resulting in higher efficiency scores of all remaining units
2. In the case of regulation: If limits on charged prices depend on efficiency scores, outlier elimination has important economic effects for suppliers and customers

- The question which has to be discussed is:
  "When must a very productive unit not be compared with the remaining less productive units because a comparison is not sensible?"

Table of Contents

- In the general (mainly statistical) literature on outlier detection we find three different concepts of outlier definition (See e.g. Hadi et al. (2009), Zimek et al. (2012))

1. Does the observation belong to the same distribution?
2. Is the observation too far away from the expected value?
3. Does an observation effect the result of the analysis too strongly?

Robust statistics

- Usual concepts like the z-score are not sensible in this context (Rousseeuw and Hubert (2011))
- Often used measures (mean, standard deviation) have low breakdown values
- E.g. mean and standard deviation can be carried arbitrarily far away by one single observation
- With large n the breakdown value is 0%
- $z-$score (breakdown value 0%)

$$z = \frac{x_i - \bar{x}}{\sigma_x}$$

Robust statistics

- Robust statitistics have large breakdown values, e.g. the median can resist up to 50% outliers (breakdown value is about 50%)
- The median of the absolute deviations from the median is a robust measure of dispersion (breakdown value 50%)

$$MAD = 1.483 \; median_{i=1,\ldots,n} \, |x_i - median_{j=1,\ldots,n}(x_j)|$$

- Robust score

$$r = \frac{|x_i - median_{j=1,\ldots,n}(x_j)|}{MAD}$$

- Tukey's rule: outlier outside the interval

$$[Q_1 - 1.5 \; IQR; \; Q_3 + 1.5 \; IQR]$$

Robust statistics

- Numerical example (taken from Rousseeuw and Hubert (2011))

$$6.27, \quad 6.34, \quad 6.25, \quad 63.1, \quad 6.28$$

- $z-$score

$$z = 1.79$$

- Robust score

$$r = 1277.5$$

- Tukey's rule: $[6.165;\ 6.445]$
- For usual limits (e.g. 2.576 for $\alpha = 0.01$) only robust socre (r) and Tukey's rule detect the outlier (63.1)

Robust statistics

- BUT methodologies 1 (other distribution) and 2 (distance to mean) are rather irrelevant:

1. There is no serious argument that the observed inputs/outputs are realizations of a random variable with a certain theoretical distribution.
2. The considerations focus on univariate distributions. In the present context, however, multidimensional distributions are more relevant.

   $\longrightarrow$ The most relevant question here is the extent to which the results of the analysis are influenced by individual observations.

   $\longrightarrow$ Univariate outlier analysis essentially only has the task of pointing out potential data errors

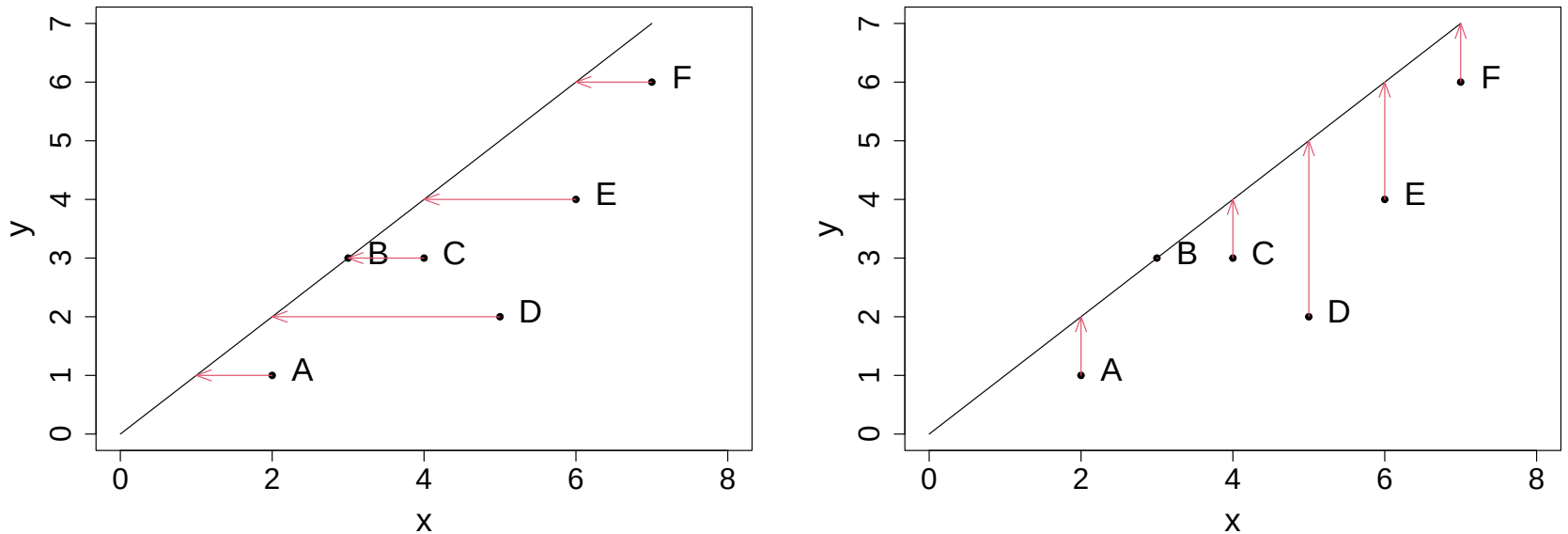
- We will focus on the Data Envelopment Analysis (DEA) in the following.

Table of Contents

- Popular method to obtain efficiency scores for firms
- Firms under analysis are compared to the most efficient "firm"
- Most often (with non-constant returns to scale) a synthetic firm obtained as a linear combination of reference firms
- Non-parametric, hence no assumptions on functional relations between inputs and outputs and data generating processes
- "Solomonic solution" in the absence of prices: each firm can "choose" the prices that maximize its own efficiency (subject to restrictions)

One input - one output case

- Most simple case, 1 input (x), 1 output (y)
- Constant returns to scale (each firm can be scaled up and down as desired)



(a) Input inefficiencies.

(b) Output inefficiencies.

Figure 1: Inefficiencies.

One input - one output case

Jacknife-idea: Leaving one firm out of the sample

- Due to the assumption of constant returns to scale (CRS) in this simple 1-input-1-output case, only efficient unit B has an impact on the remaining scores
- What is the effect of removing B from the sample?
- After removing B the less efficient firm F is the new benchmark
- The change in efficiency is identical for all units: 16.7%
- In this simple case the change equals: productivity of B / productivity F

Two input - two output case

- We now consider 10 units and two inputs and two outputs
- For all 10 units we calculate the average change in the remaining 9 units efficiency score when this unit is left out

Table 1: long.

Var./Res.	1	2	3	4	5	6	7	8	9	10
x1	2.00	3.00	4.00	5.00	6.00	7.00	8.00	4.00	7.00	6.00
x2	2.00	4.00	3.00	6.00	5.00	6.00	3.00	9.00	6.00	7.00
y1	1.00	3.00	3.00	8.00	4.00	6.00	2.00	1.00	9.00	7.00
y2	2.00	1.00	7.00	3.00	5.00	5.00	4.00	4.00	3.00	7.00
θ	0.61	0.63	1.00	1.00	0.67	0.76	0.63	0.57	1.00	0.99
%	0.00	0.00	25.71	3.19	0.00	0.00	0.00	0.00	1.19	0.00

- In the example we find that
  - Unit 3 is efficient, disregarding this unit increases on average the score by 26%
  - Unit 4 is efficient, disregarding this unit increases on average the score by 3%
  - Unit 9 is efficient, disregarding this unit increases on average the score by 1%

Two input - two output cas

- Due to the assumption of CRS only few units are efficient
- Only these units influence the score of the remaining units
- We obtain the average change of scores for other units "caused" by this influential unit
- Hence, we obtain a ranking of the most influencial units (3, 4, 9)

The problem of masking

- The term "mask" means that an outlier hides other outliers (Simar (2003), Clermont and Schaefer (2019)).
- If the worst outlier would be removed, we might find a second outlier and so on
- Implies an iterative procedure $\longrightarrow$ Leave 1 out, leave 1 out, leave 1 out, ...
- Alternative: leave several observations out
- $\longrightarrow$ Leave 1 out, leave 2 out, leave 3 out, ...
- Problem: Becomes rapidly computer intensive

The problem of masking

- Alternative is (Khezrimotlagh et al. (2020)):

1. to start with elimination of a larger set of observations (e.g.with highest scores) and
2. to add individual observations
3. stop adding observations if a "significant" change in the score distribution is observed

- Suggested criterion: outlier if maximal difference in distribution functions of scores (with/without potential outlier) is significant (Kolmogorov)

Table of Contents

- In this context there are only two relevant occurences of outliers
  1. Measurement errors
  2. Mixing non-homogeneous DMUs (non-comparable production environments)
- Standard routines may help to flag observations that might contain measurement errors. Decision after inspection of individual observations
- It is a purely substantive (and not a statistical!) question whether the production circumstances of influential units are comparable or not
- Beeing very efficient is by no means for itself a reason to be regarded as an outlier!

In a recent overview article Khezrimotlagh et al. (2020) conclude:

> "There are not yet any known methods which can be applied without careful and essential user judgment to deal with presence of outliers."
>
> "Finally, the most important step is that users should carefully investigate each flagged observation to determine whether or not such a flagged observation is an outlier."
>
> "... none of the available procedures are dependable to estimate a production frontier without incorporating a significant amount of user judgment."

References

Marcel Clermont and Julia Schaefer, Identification of outliers in data envelopment analysis: An approach using structure-detecting statistical procedures, *Schmalenbach Business Review*, 71 (4):475–496, 2019.

Ali S Hadi, A. H. M. Rahmatullah Imon, and Mark Werner, Detection of outliers, *Wiley Interdisciplinary Reviews: Computational Statistics*, 1(1):57–70, 2009.

Dariush Khezrimotlagh, Wade D Cook, and Joe Zhu, A nonparametric framework to detect outliers in estimating production frontiers, *European Journal of Operational Research*, 286(1): 375–388, 2020.

Peter J Rousseeuw and Mia Hubert, Robust statistics for outlier detection, *Wiley interdisciplinary reviews: Data mining and knowledge discovery*, 1(1):73–79, 2011.

Léopold Simar, Detecting outliers in frontier models: A simple approach, *Journal of Productivity Analysis*, 20:391–424, 2003.

Arthur Zimek, Erich Schubert, and Hans-Peter Kriegel, A survey on unsupervised outlier detection in high-dimensional numerical data, *Statistical Analysis and Data Mining: The ASA Data Science Journal*, 5(5):363–387, 2012.