Robust Statistics summary

A. Tsanakas, Prof. Ioannis Pitas
Aristotle University of Thessaloniki
pitas@csd.auth.gr
www.aiia.csd.auth.gr
Version 2.5

Robust Statistics

- Outliers
- Measures of Robustness
  - Sensitivity Curve (SC)
  - The Influence Function
  - Breaking Point
- Robust Estimators
  - L - Estimators
  - M - Estimators
  - S - Estimators

Robust Statistics

- In parametric statistics, a priori assumptions such as regularity, independence and linearity play a big role.

- In many statistical algorithms, small deviations from the assumptions can lead to large errors in the result.

- These small deviations are due to the presence of ***general errors*** (*outliers*).

Robust Statistics

- The problem we described before is being solved by ***robust statistics*** which are an extension of parametric statistics.

- The term robustness means insensitive to small deviations from the model's assumptions.

- Also, robustness theories are defined as deviations from the assumptions of parametric models.

Statistics

- Outliers
- **Measures of Robustness**
  - **Sensitivity Curve (SC)**
  - **The Influence Function**
  - **Breaking Point**
- **Robust Estimators**
  - **L - Estimators**
  - **M - Estimators**
  - **S - Estimators**

Outliers

- An outlier is an observation that has an unusual distance from other observations in a random sample from a population.
- This definition leaves it up to the analyst to decide what will be considered unusual.
- Many times removing an outlier is not the best solution. An outlier can give us important information about the sample.

Outliers

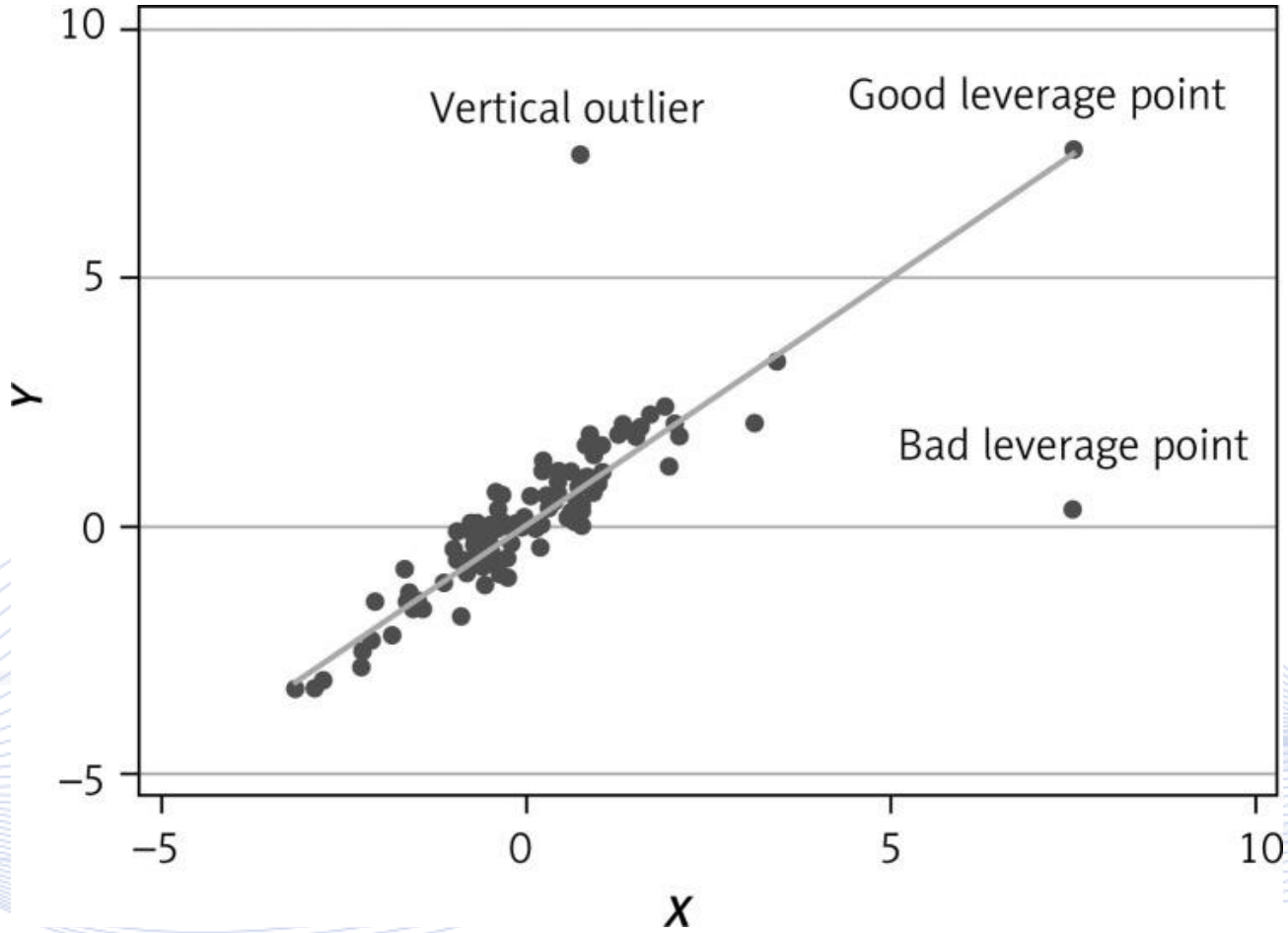
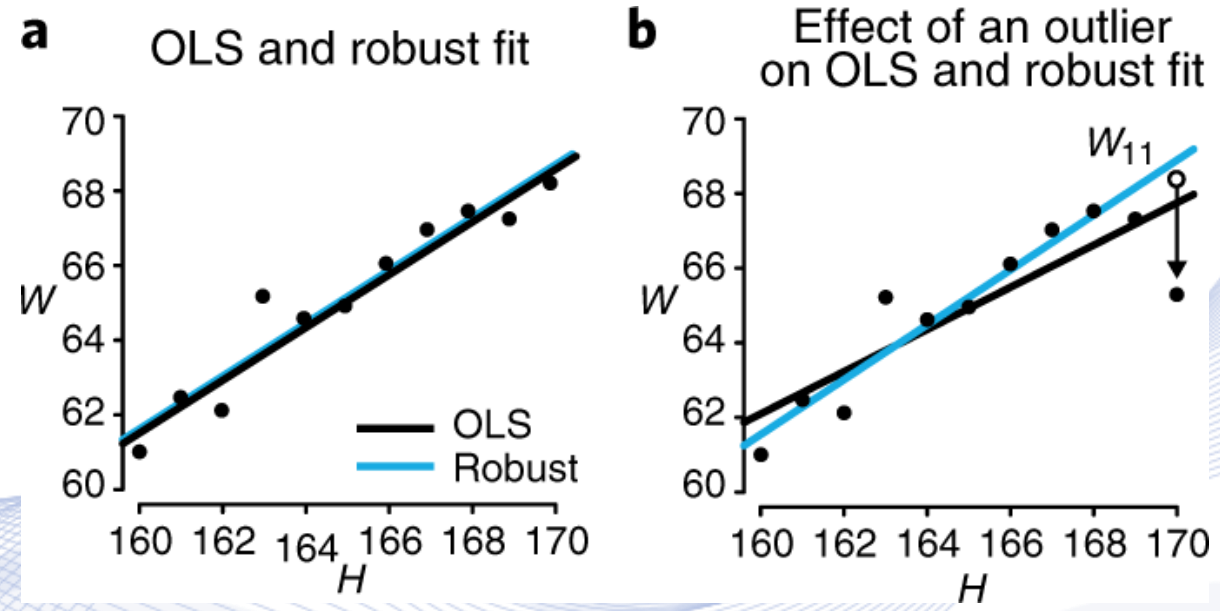


Figure: Outliers ([VAD20])

Outliers

In this figure we can observe:

1. The effect of the outliers on ***Ordinary Least Squares*** (OLS).
2. How robust statistics help us address this problem.



([GRE19])

Statistics

- **Outliers**
- **Measures of Robustness**
  - **Sensitivity Curve (SC)**
  - **The Influence Function**
  - **Breaking Point**
- **Robust Estimators**
  - **L - Estimators**
  - **M - Estimators**
  - **S - Estimators**

Measures of Robustness

To measure the robustness of an estimator, commonly use the following measures:

- ***Sensitivity Curve (SC)***
- ***The Influence Function***
- ***Breaking Point***

Statistics

- Outliers
- Measures of Robustness
  - Sensitivity Curve (SC)
  - The Influence Function
  - Breaking Point
- Robust Estimators
  - L - Estimators
  - M - Estimators
  - S - Estimators

Measures of Robustness - The Influence Function

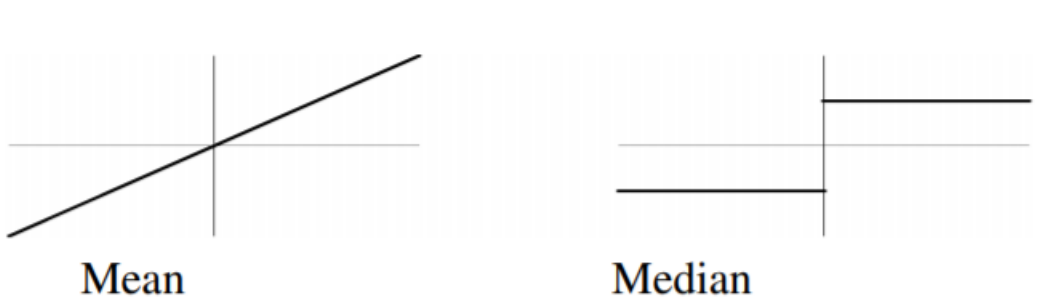
- The ***Influence Function*** describes how the estimator reacts to a small amount of contamination at any point y_0.

- We want to estimate the parameter θ of a distribution F. Let the functional $T_n = T(F_n)$ estimates θ.

Measures of Robustness - The Influence Function

- ***Example:*** For the mean $T(F) = E_F[X]$ it gives:

$$IF(x;T;F) = \lim_{t\to 0} \frac{(1-t)T(F) + tx - T(F)}{t} = x - T(F)$$



The influence Function of Mean and Median

Statistics

- Outliers
- Measures of Robustness
  - Sensitivity Curve (SC)
  - The Influence Function
  - Breaking Point
- Robust Estimators
  - L - Estimators
  - M - Estimators
  - S - Estimators

Breaking Point

- The ***breakdown point*** of an estimator is the proportion of incorrect observations an estimator can handle before giving an incorrect result.
- If replace m of observation by any outliers and estimator stays in a bounded set, but doesn't when we replace $(m+1)$, the breakdown point of the estimator at that data set is:

$$bdp = m/n$$

Statistics

- Outliers
- Measures of Robustness
  - Sensitivity Curve (SC)
  - The Influence Function
  - Breaking Point

- Robust Estimators
  - L - Estimators
  - M -Estimators
  - S - Estimators

Robust Estimators

Robust estimators are estimation techniques that are insensitive to small departures from the idealized assumptions. The main resilient appraisers are the following:

- ***L - Estimators***
- ***M - Estimators***
- ***S - Estimators***

Statistics

- Outliers
- Measures of Robustness
  - Sensitivity Curve (SC)
  - The Influence Function
  - Breaking Point
- Robust Estimators
  - L - Estimators
  - M - Estimators
  - S - Estimators

Robust Estimators – L-estimators

Also, typical examples of such estimators are the ***a-trimmed mean***:

$$L_n = \left(\frac{1}{(n-2) * [na]}\right) \sum_{i=[na]+1}^{n-[na]} X_{n:i}$$

Robust Estimators – L - Estimators

The ***a-Winsorized mean***:

$$L_n = \frac{1}{n}\{[na]X_{n:[na]} + \sum_{i=[na]+1}^{n-[na]} X_{n:i} + [na] + X_{n:n-[na]+1}\}$$

where $0 < a < 1/2$ and $[x]$ is the largest integer k satisfying $k \leq x$.

Statistics

- Outliers
- Measures of Robustness
  - Sensitivity Curve (SC)
  - The Influence Function
  - Breaking Point
- Robust Estimators
  - L - Estimators
  - M - Estimators
  - S - Estimators

Robust Estimators – M - Estimators

- ***M - Estimators*** are a generalization of ***Maximum Likelihood Estimation*** (MLE).

- We minimize, instead of $\log f(x, \theta)$ as in MLE, a more general function $\rho(x, \theta)$.

Statistics

- Outliers
- Measures of Robustness
  - Sensitivity Curve (SC)
  - The Influence Function
  - Breaking Point
- Robust Estimators
  - L - Estimators
  - M - Estimators
  - S - Estimators

Robust Estimators – S - Estimators

- The aim of ***S - Estimators*** is to have a simple high-breakdown regression estimator, which share the flexibility and nice asymptotic properties of M - Estimators.

Bibliography

[PIT1990] I. Pitas and A. N. Venetsanopoulos, Nonlinear digital filters: principles and applications, Kluwer Academic, 1990.

[ALA18] Almetwally, E., and H. Almongy. "Comparison between m estimation, s estimation, and mm estimation methods of robust estimation with application and simulation." *International Journal of Mathematical Archive* 9.11 (2018): 1-9.

[RUI12] Ruiz-Gazen, Anne. "Robust statistics: a functional approach." *Annales de l'ISUP*. Vol. 56. 2012.

[WAT04] Watkins, Joseph C. "An introduction to the science of statistics: From theory to implementation." (2019).

[HUB04] Huber, Peter J. *Robust statistics*. Vol. 523. John Wiley & Sons, 2004.

[ROM11] Rousseeuw, Peter J., and Mia Hubert. "Robust statistics for outlier detection." *Wiley interdisciplinary reviews: Data mining and knowledge discovery* 1.1 (2011): 73-79.

[HAM01] Hampel, Frank R. "Robust statistics: A brief introduction and overview." *Research report/Seminar für Statistik, Eidgenössische Technische Hochschule (ETH)*. Vol. 94. Seminar für Statistik, Eidgenössische Technische Hochschule, 2001.

Bibliography

[VAD20] Varin, Sacha, and Demosthenes B. Panagiotakos. "A review of robust regression in biomedical science research." *Archives of Medical Science: AMS* 16.5 (2020): 1267.

https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7444710/

[GRE19] Greco, Luca, et al. "Analyzing outliers: robust methods to the rescue." *Nature methods* 16.4 (2019): 275-7.

[JUR84] Jurečková, Jana. "21 M-, L-and R-estimators." Handbook of statistics 4 (1984): 463-485.

Q & A

Thank you very much for your attention!

More material in
http://icarus.csd.auth.gr/cvml-web-lecture-series/

Contact: Prof. I. Pitas
pitas@csd.auth.gr