Comparing Robust Estimators with an Application to Environmental Data

Masuma Mannan, Nubaira Rizvi, Dr. Evrim Oral.

LSU Health NEW ORLEANS School of Public Health

Louisiana State University Health Sciences Center School of Public Health, Biostatistics and Data Science Program

publichealth.lsuhsc.edu

Background

Robust statistics aims to create methods less affected by outliers or model assumption violations.

They offer more reliable results compared to classical methods, especially in the presence of outliers.

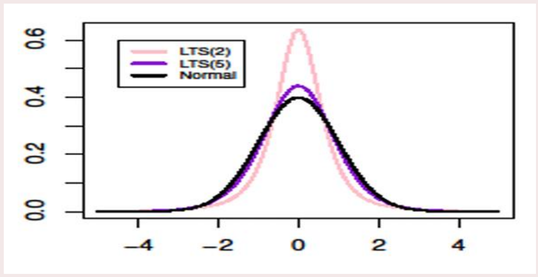
Sanaullah et al. (2019) and Ahmad et al. (2023) used a Best Linear Unbiased (BLUE) type estimator based on order statistics to introduce novel ratio-type estimators.

Although they examined the properties of this estimator within the survey-sampling framework, the robustness of BLUE-type estimators needs further investigation.

Objectives

- Evaluate the robustness properties of the BLUE-type location estimator and compare it with other established robust estimators, such as Tiku's modified maximum likelihood estimator (MMLE).
- Demonstrate the performance of these estimators by analyzing environmental data.

Long Tail Distributions



- A long-tailed symmetric distribution is a statistical distribution characterized by gradual declines in the tails (extreme values).

$$f(x)=\frac{1}{\sigma k^{1/2}\beta(\frac{1}{2},p-\frac{1}{2})}\left\{1+\frac{(x-\mu)^2}{k\sigma^2}\right\}^{-p},\quad -\infty<x<\infty,$$

- Long-tailed symmetric distributions have tails that gradually taper off, extending far from the central point with a symmetrical pattern.

Methods

We assumed that the underlying distribution of interest follows a long-tailed symmetric (LTS) family distribution. Utilizing the LTS distribution, we theoretically derived the BLUE-type location estimator

$$\tilde{\mu} = [(w'\Omega^{-1}w)^{-1}]'\,(w'\Omega^{-1}y)$$

Methods (Cont.)

We then empirically compared its robustness properties with Tiku's robust MMLE

$$\hat{\mu}=\left\{\sum_{i=1}^{n}\beta_i x_{(i)}\right\}\Big/ m \quad \left(m=\sum_{i=1}^{n}\beta_i\right)$$

through an extensive simulation study which was conducted using R.

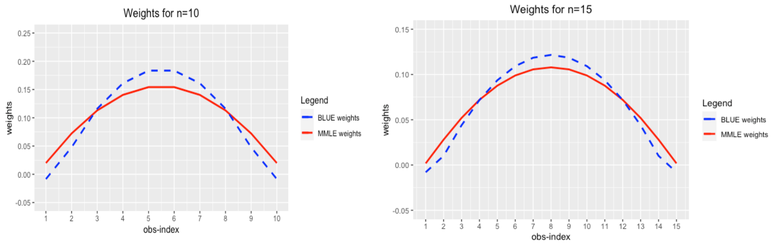
Simulation Settings

We considered four different model settings in our simulation study:

1. True Model
2. Outlier Model
3. Mixture Model (0.90 LTS + 0.10 Laplace)
4. Contaminated Model (0.90 LTS + 0.10 LTS (a=0,b=c))

The sample size is 10 and 15, p = 2.5 and
c = 2, 3, 4.

Robustness



Sample		True	Mixture	Outlier	Contaminated
n=10	R.E	1.02	1.05	1.17	1.22
	Kurtosis	2.95	3.10	3.57	3.60
n=15	R.E	1.01	1.03	1.06	1.11
	Kurtosis	3.50	3.73	4.44	4.80

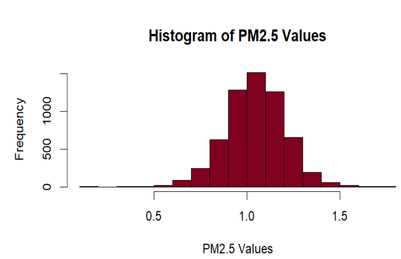
- R.E = MSE(MMLE) / MSE(BLUE)
- Kurtosis (>3) indicates that the distribution is peaked and possesses thick tails.

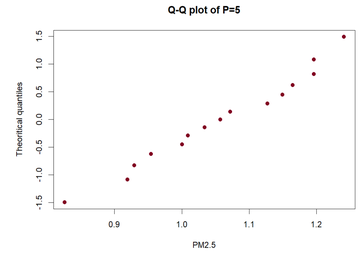
Results

- In comparison to the MML estimator, the BLUE-type estimator of the mean have higher efficiency when the tail of the distribution gets thicker.
- As the sample size increases, they both have similar efficiency for all four models (not shown here)

Application: PM 2.5 Data

- The outdoor air quality data 2023 contains PM2.5 (particulate matter with a diameter of 2.5 micrometers or less) monitoring conducted by the Environmental Protection Agency (EPA) in Louisiana.
- The variable PM2.5 comes from a LTS distribution with P=5 as shown by the Q-Q plot.
- The three estimation techniques are applied to estimate the parameters.





Estimators	$\bar{X}$	MMLE	BLUE
Mean	1.05	1.07	1.09
SD	0.120	0.13	0.04

Conclusion

Overall, the BLUE-type estimator have higher efficiency than MML estimators when n is small.

For the tails of the LTS, the weights of the BLUE-type estimator are smaller than the weights of the MMLE, making it more robust when n is small

References

1. M.L. Tiku and R.P. Suresh, "A new method of estimation for location and scale parameters", J.Stat. Plann. Inference, vol. 30,no. 2, pp. 281-292, Feb. 1992.
2. E. H. Lloyd, "Least Squares Estimation of Location and Scale Parameters Using Order Statistics", Biometrika, vol. 39, no. 1/2, pp. 88 95, Apr. 1952.