Musik in bayern
ISSN: 0937-583x Volume 89, Issue 12 (Dec -2024)
https://musikinbayern.com DOI https://doi.org/10.15463/gfbm-mib-2024-*358*

"Improving Reliability in Data Analysis: Techniques in Robust Statistics"

First Author
Mrs. P. Matlida Shanthini,
Assistant Professor,
Department of Mathematics, SRM Institute and Technology, Faculty of Science and Humanities, Ramapuram, Chennai. Matildap@srmist.edu.in
Corresponding Author
Dr. B.Sumithra.,
M.Sc.,M.Phil.,Ph.D (stat)., PGDCA., DCST, Assistant Professor,
Department of Mathematics, SRM Institute and Technology, Faculty of Science and Humanities, Ramapuram, Chennai. sumithra.b82@gmail.com

To Cite this Article

Mrs. P. Matlida Shanthini, **Corresponding Author** Dr. B.Sumithra**" Improving Reliability in Data Analysis: Techniques in Robust Statistics"** *Musik In Bayern, Vol. 89, Issue 12, Dec 2024, pp64-72*

Article Info

Received: 14-10-2024 Revised: 12-11-2024 Accepted: 22-11-2024 Published: 07-12-2024

Abstract

Robust statistics is a branch of statistics that develops methods to provide reliable and effective results even when data deviates from the assumptions of classical techniques. This includes handling issues such as outliers, non-normal distributions, and model misspecifications, which commonly arise in real-world datasets. Key principles of robust statistics include resistance to extreme values, high breakdown points, and efficiency in ideal conditions. Common robust methods include measures like the median and interquartile range, trimmed means, robust regression techniques such as M-estimators, and multivariate techniques like the Minimum Covariance Determinant (MCD) estimator. These methods are widely applied in fields such as finance, medicine, environmental science, and machine learning, where data contamination and model reliability are critical concerns. This paper explores the concepts, methodologies, and applications of robust statistics, emphasizing their importance in modern data analysis.

Keyword: Robust Statistics, M-estimators, trimmed means, robust regression and Minimum Covariance Determinant

Introduction to Robust Statistics

ISSN: 0937-583x Volume 89, Issue 12 (Dec -2024)
https://musikinbayern.com DOI https://doi.org/10.15463/gfbm-mib-2024-*358*
Robust statistics is a bough of statistics that focuses on developing methods and techniques that remain reliable and effective even when the assumptions underlying standard statistical methods are violated. Unlike classical statistical move towards, which can be sensitive to outliers or deviations from model assumptions (such as normality), robust statistical methods aim to create results that are less influenced by such irregularities.

1. **Sensitivity to Outliers**: Classical methods, such as the mean or ordinary least squares (OLS) regression, can be a great deal influenced by outliers, leading to misleading conclusions.
2. **Non-Normal Distributions**: Many real-world datasets stray from idealized assumptions like normality or homoscedasticity.
3. **Model Misspecification**: Robust methods help alleviate the impact of small violations in model assumptions; ensuring results remain interpretable and useful.

Principles of Robust Statistics

1. **Resistance**: Robust methods limit the influence of extreme values or outliers.
2. **Breakdown Point**: This measures the proportion of contaminated data a method can handle before producing incorrect results. Higher breakdown points indicate more robust methods.
3. **Efficiency**: Robust methods aim to perform well under ideal conditions while maintaining reliability under non-ideal conditions.

Common Robust Statistical Methods

1. **Robust Measures of Central Tendency**:
    - **Median**: A robust alternative to the mean, as it is unaffected by extreme values.
    - **Trimmed Mean**: Calculated by discarding a fixed percentage of extreme values at both ends of the dataset before computing the mean.
2. **Robust Measures of Spread**:
    - **Interquartile Range (IQR)**: The range between the 25th and 75th percentiles, robust to outliers.
    - **Median Absolute Deviation (MAD)**: The median of the absolute deviations from the dataset's median.
3. **Robust Regression**:
    - **Least Median of Squares (LMS)**: Minimizes the median of squared residuals, making it robust to outliers.
    - **M-Estimators**: A flexible framework for regression, generalizing least squares by weighting residuals to reduce the impact of outliers.
4. **Nonparametric Methods**:

ISSN: 0937-583x Volume 89, Issue 12 (Dec -2024)
https://musikinbayern.com DOI https://doi.org/10.15463/gfbm-mib-2024-*358*

- These methods, such as rank-based tests (e.g., Wilcoxon or Kruskal-Wallis tests), make minimal assumptions about the data distribution, enhancing robustness.

Applications of Robust Statistics

Robust statistics are used in fields such as:

- Finance: Handling noisy or erroneous financial data.
- Medicine: Analyzing clinical trial results with potential measurement errors.
- Environmental Science: Managing outliers in meteorological or ecological datasets.
- Machine Learning: Robust loss functions to mitigate the impact of noisy labels or outliers.

Advantages and Limitations

Advantages:

- Improved reliability in real-world data.
- Protection against data contamination and model misspecification.
- Often more interpretable results when data deviates from assumptions.

Limitations:

- Can be less efficient than classical methods under ideal conditions.
- May involve more computational complexity.
- Choice of robust methods can depend on specific assumptions about data contamination.

Python Code with Examples

Median:

```
set.seed(42)

data <- c(rnorm(100, mean = 50, sd = 5), 200)  # Outlier included

# Median

median(data)
```

50.44916

ISSN: 0937-583x Volume 89, Issue 12 (Dec -2024)
https://musikinbayern.com DOI https://doi.org/10.15463/gfbm-mib-2024-*358*
The median is the middle value of a dataset when the data is sorted in ascending order. Unlike the mean, the median is a robust measure of central tendency, meaning it is not significantly affected by outliers. The median is the middle value of a dataset when the data is sorted in ascending order. Unlike the mean, the median is a robust measure of central tendency, meaning it is not significantly affected by outliers.

Mean:

A **trimmed mean** is a measure of central tendency that removes a specified proportion of extreme values (from both ends) in a dataset before calculating the mean. This method helps reduce the impact of outliers and provides a more robust measure of the "typical" value in the data.

Trimmed mean (remove extreme 10% from each end)

mean(data, trim = 0.1)

50.50601

Interpretation:

- After trimming the extreme 10% from each end, the calculated trimmed mean is **50.50601**.
- This value represents the average of the middle 80% of your data, providing a central tendency less influenced by extreme values.

Regression:

Formula for regression:

The model being fit is $y = \beta_0 + \beta_1 x + \varepsilon$,

where:

- y is the dependent variable.
- x is the independent variable.
- β_0 is the intercept.
- $\beta_1\ x$ is the slope (coefficient of x).
- ε represents the residuals.

Robust Regression:

- The function **`rlm()`** uses an iterative process to reduce the influence of outliers.

ISSN: 0937-583x Volume 89, Issue 12 (Dec -2024)
https://musikinbayern.com DOI https://doi.org/10.15463/gfbm-mib-2024-*358*

- It minimizes a robust measure of scale, often using a weighting scheme that downweights observations with large residuals.

```
# Install and load the MASS package

if (!require("MASS")) install.packages("MASS")

library(MASS)

# Sample data

x <- 1:100

y <- 2 * x + rnorm(100, sd = 5)

y[95:100] <- y[95:100] + 50  # Add outliers
```

Call: rlm(formula = y ~ x)

Residuals:

Min	1Q	Median	3Q	Max
-9.3827	-2.5914	-0.1015	2.5648	57.3792

Coefficients	Value	Std. Error	t value
Intercept (β0)	-2.2612	1.0744	-2.1047
Slope (β1)	2.0400	0.0185	110.4469

Residual standard error: 3.826 on 98 degrees of freedom

The above table shows that the intercept is2.2612 and Slope is 2.0400. Outliers have minimal influence on the regression line.

```
abline(robust_model, col = "red", lwd = 2)

if (!require("robustbase")) install.packages("robustbase")

library(robustbase)

# Robust covariance matrix
```

ISSN: 0937-583x Volume 89, Issue 12 (Dec -2024)
https://musikinbayern.com DOI https://doi.org/10.15463/gfbm-mib-2024-*358*
covMcd(data.frame(x, y))

```
Minimum Covariance Determinant (MCD) estimator approximation.
Method: Fast MCD(alpha=0.5 ==> h=51); nsamp = 500; (n,k)mini = (3
00,5)
Call:
covMcd(x = data.frame(x, y))
Log(Det.):  7.395

Robust Estimate of Location:
    x      y
47.27  93.80
Robust Estimate of Covariance:
        x     y
x   768.4  1561
y  1560.7  3184
```

Minimum Covariance Determinant (MCD) **analysis, a robust statistical method for estimating the mean (location) and covariance of multivariate data while minimizing the influence of outliers.** MCD Overview**:**

The MCD method identifies a subset of the data (called "h-subset") with $h = floor(n.\alpha)$ observations that minimizes the determinant of the covariance matrix, providing robust estimates of location and covariance.

$\alpha = 0.5,$ Indicates that 50% of the data points are used to calculate the robust estimates.

$h = 51,$ The size of the subset used for robust estimation in your case.

The log determinant ($\log(\mathrm{Det}(\Sigma))=7.395$) reflects the compactness of the robustly estimated data subset. A smaller value indicates tighter clustering of the selected subset.

Robust Estimate of Location:

- Location estimates represent the robust mean values of the variables x and y, calculated from the selected h-subset.

  Mean(x) = 47.27

  Mean(y) = 93.80

ISSN: 0937-583x Volume 89, Issue 12 (Dec -2024)
https://musikinbayern.com DOI https://doi.org/10.15463/gfbm-mib-2024-*358*
These values indicate the central tendency of the data, with outliers excluded.

The positive robust covariance (1560.7) implies a strong positive relationship between x and y, but robustly estimated to minimize the impact of outliers.

Applications

- The MCD estimates are useful in detecting outliers or high-leverage points, assessing data structure robustly, and preparing for downstream multivariate analyses (e.g., robust PCA).
- Robust Analysis of Variance (ANOVA) methods are alternatives to classical ANOVA, designed to handle violations of key assumptions such as normality and homogeneity of variance. These robust methods are especially useful in datasets where extreme values, unequal variances, or non-normal distributions might otherwise distort results.
- One commonly used robust ANOVA method is the **trimmed means one-way ANOVA** (as seen in the output provided). This method trims a percentage of the largest and smallest values in each group before computing the test statistic, which minimizes the influence of outliers.

Robust Analysis of Variance (ANOVA) methods are alternatives to classical ANOVA, designed to handle violations of key assumptions such as normality and homogeneity of variance. These robust methods are especially useful in datasets where extreme values, unequal variances, or non-normal distributions might otherwise distort results.

One commonly used robust ANOVA method is the **trimmed means one-way ANOVA** (as seen in the output provided). This method trims a percentage of the largest and smallest values in each group before computing the test statistic, which minimizes the influence of outliers.

```
# Robust ANOVA

t1way(values ~ group)
```

```
t1way(formula = values ~ group)

Test statistic: F = 94.049
Degrees of freedom 1: 2
Degrees of freedom 2: 21.61
p-value: 0

Explanatory measure of effect size: 0.84
Bootstrap CI: [0.74; 0.94]
```

ISSN: 0937-583x Volume 89, Issue 12 (Dec -2024)
https://musikinbayern.com DOI https://doi.org/10.15463/gfbm-mib-2024-*358*
The robust ANOVA results demonstrate a significant effect of the grouping variable on the outco me variable, with a large effect size (R^2=0.84) and strong statistical evidence ($p < 0.001$). The bo otstrap confidence interval further supports the reliability of these findings, ensuring robustness t o potential violations of assumptions.

Introduction to Multivariate Outlier Detection Using Mahalanobis Distance

In multivariate analysis, detecting outliers is critical because outliers can significantly affect statistical models, leading to misleading conclusions. One commonly used technique for detecting outliers in multivariate data is the **Mahalanobis distance**.

The Mahalanobis distance is a measure of the distance between a point and the center of a multivariate distribution, considering the data's covariance structure. It identifies points that deviate significantly from the central tendency while accounting for the correlations between variables.

```
# Mahalanobis Distance for multivariate data

data_multivariate <- data.frame(x = rnorm(100), y = rnorm(100))

mahal_dist <- covMcd(data_multivariate)$mah

outliers <- which(mahal_dist > quantile(mahal_dist, 0.975))

# Outliers

print(outliers)
```

```
8 53 78
```

The Mahalanobis distance effectively flagged three observations (**indices 8, 53, and 78**) as potential outliers. These data points are unusually far from the multivariate center of the dataset, considering the covariance structure of the variables. Detecting such outliers is important for ensuring the integrity and reliability of subsequent analyses.

ISSN: 0937-583x Volume 89, Issue 12 (Dec -2024)
https://musikinbayern.com DOI https://doi.org/10.15463/gfbm-mib-2024-*358*

Conclusion

Robust statistics plays an indispensable role in modern data analysis by addressing real-world data irregularities such as outliers, non-normality, and model violations. Its methods ensure reliability, making it a valuable tool across diverse fields like finance, medicine, and machine learning. While computational demands and efficiency in ideal conditions are considerations, the benefits of robust methods in contaminated datasets outweigh these limitations.

REFERENCE

1. Maronna, R. A., Martin, R. D., & Yohai, V. J. (2006).*Robust Statistics: Theory and Methods.* Wiley.
2. Huber, P. J., & Ronchetti, E. M. (2009).*Robust Statistics.* Wiley.
3. Hampel, F. R., Ronchetti, E. M., Rousseeuw, P. J., & Stahel, W. A. (1986). *Robust Statistics: The Approach Based on Influence Functions.* Wiley.4.
4. Rousseeuw, P. J. (1984)."Least Median of Squares Regression." *Journal of the American Statistical Association, 79*(388), 871-880.
5. Rousseeuw, P. J., & Leroy, A. M. (1987)."Robust Regression and Outlier Detection."
6. Tukey, J. W. (1977)."Exploratory Data Analysis." Addison-Wesley.
7. Van der Vaart, A. W. (2000)."Asymptotic Statistics." Cambridge University Press.