Al Kut Journal of Economics and Administrative Sciences /ISSN: 1999 -558X /ISSN Online 2707-4560/ Vol (16) Issue: 53-2024 (August)
A special issue of the proceedings of the Fourth International Conference of the Faculty of Administration and Economics, University of Wasit

AL KUT JOURNAL OF ECONOMIC AND ADMINISTRATIVE SCIENCES
Publisher: College of Economics and Management - Wasit University

Estimating a limited population mean for discrete linear regression using some robust regression methods

Mustafa Habib Mahdi

Middle Technical University , Institute of Administration/ Al-Rusaffa , Department of Statistics Techniques.

Baghdad, Iraq.

mustafa_habib@mtu.edu.iq

http://dx.doi.org/10.29124/kjeas.1653.11

Abstract :-

This study concentrates on estimating the population mean within stratified random sampling utilizing discrete regression estimators, juxtaposed against classical estimators relying on mean squared error (MSE) and efficiency criterion (RE). Despite the ordinary least squares (OLS) estimator's efficacy in accounting for error covariance, it exhibits heightened sensitivity towards outliers. To tackle this challenge, robust regression estimators, such as LTS, Tukey M, and Hampel M, were implemented alongside robust covariance and covariance matrices (MCD and MVE). This study was applied to real data from the Central Bureau of Statistics, specifically the results of the hotel survey in Dohuk Governorate for the year 2020, which included a group of hotels with a size of 115 hotels. By describing the data represented in the total wages paid, the number of employees, the results showed that the (LTS) estimator achieved high efficiency in estimating the regression parameter compared to the rest of the estimators used in the study.

Keywords: Discrete regression estimator ($\bar{\mathbf{y}}_{\mathbf{lrs}}$), outliers, robust regression methods, efficiency, stratified random sampling.

1. Introduction

The discrete regression estimator is a common method used in stratified random sampling to estimate the population parameter. This estimator is based on fitting individual regression models within each stratum, and then using these models to estimate the population mean. By estimating the regression parameter independently in each layer, this approach provides greater flexibility in capturing the relationship between the response and predictor variable in each layer.

The discrete Regression Estimator is particularly effective when dealing with dispersion in the population, where the relationships between variables of interest change across different strata. By allowing for discrete models within each layer, this estimator can keep up with the unique properties and correlations present in each layer. As a result, it provides more accurate and reliable estimates.

M-estimation refers to a method of estimating parameters of a statistical model. It involves minimizing a criterion function, often denoted as ρ, which depends on a set of sample data. Different choices of the function ρ can lead to different estimators.

Hampel (1971) and Tukey (1977) are prominent figures in the development of robust statistics. They proposed various robust estimators and suggested different ρ functions to be used in M-estimation. These ρ functions are designed to be less sensitive to outliers and deviations from model assumptions compared to traditional estimation methods like maximum likelihood estimation (MLE). By choosing appropriate ρ functions, one can obtain robust estimators that perform well even in the presence of outliers or other departures from standard assumptions. In LTS (Least Trimmed Squares) method, the squared errors are sorted and OLS is done by using observations based on the first (smallest) z errors. So, the results are not affected by outliers (Li, 2004).The study aims to estimate the linear regression parameter for predicting the population mean using various discrete regression methods, specifically robust techniques like LTS (Least Trimmed Squares), Hampel M , and Tukey M. Additionally, it incorporates robust variance and covariance matrix estimations such as MVE (Minimum Volume Ellipsoid) and MCD (Minimum Covariance Determinant). These methods are implemented within the framework of stratified random sampling to attenuate the influence of outliers and bolster the efficacy of these techniques. Subsequently, a comparison is drawn between the estimators generated by these robust methods and the traditional Ordinary Least Squares (OLS) method to ascertain the optimal estimator based on the Mean Squared Error (MSE) criterion.

2. Stratified Random Sampling Design (STRS)

When dealing with heterogeneous population units, it's crucial to improve the accuracy of estimates obtained from a simple random sample (SRS) by addressing this diversity. One effective strategy involves dividing the population into homogeneous subgroups called strata. Each stratum exhibits internal homogeneity and differs from the others. After determining the number of strata, samples are drawn independently and randomly from each stratum. This ensures that the selection process is uniform across all strata, helping to mitigate the impact of population heterogeneity on the estimation process. (William G. Cochran , 1977)

3. Linear Regression Estimators to Estimate the population Mean in Stratified Random Sampling

The estimate for the population mean in stratified random sampling, obtained through the discrete regression estimator ($\bar{y}_{lrs}$) , depends on the linear regression of Y_i over X_i :

3.1 Discrete regression Estimator ($\bar{y}_{lrs}$)

In this estimator, the regression method is utilized separately for simple random samples within each stratum. This procedure entails estimating a set of regression coefficients, denoted as b_h, for each individual stratum. These regression estimates for each stratum are subsequently aggregated by computing a weighted average, with the relative sizes of the strata serving as the weighting factors. Precisely, the weights are determined based on the proportional sizes of the strata. (Dick J. Brus , 2022)

This type of estimation is employed when there is sufficient evidence to posit that the actual values

(B_h) within each stratum are homogeneous or similar to one another.

We calculate the regression estimate for the average of each stratum, namely. (Cochran , 1977)

$$\bar{y}_{lrh} = \bar{y}_h + b_h \left(\bar{X}_h - \bar{x}_h \right) \qquad (1)$$

In contrast, the separate regression estimator's parameter, denoted as b_h, is estimated using the ordinary least squares (OLS) method .

$$b_h = \frac{s_{xyh}}{s_{xh}^2}$$

s_{xh}^2 is the sample variance of the auxiliary variable in layer h .

$$s_{xh}^2 = \frac{\sum_{i=1}^{n_h} (x_{hi} - \bar{x}_h)^2}{n_h}$$

s_{xyh} is the covariance between the auxiliary variable and the study variable in class h.

$$s_{xyh} = \frac{\sum_{i=1}^{n_h} (x_{hi} - \bar{x}_h)(y_{hi} - \bar{y}_h)}{n_h - 1}$$

Stratified sample averages for the two variables (x and y).

$$\bar{y}_h = \frac{\sum_{i=1}^{n_h} y_{hi}}{n_h}$$

$$\bar{x}_h = \frac{\sum_{i=1}^{n_h} x_{hi}}{n_h}$$

The stratified community mean for the variable X.

$$\bar{X}_h = \frac{\sum_{i=1}^{N_h} X_{hi}}{N_h}$$

Stratum weight for layer h.

$$W_h = \frac{N_h}{N}$$

Then the mean of the discrete stratified regression is:

$$\hat{\bar{Y}}_{lrs} = \sum_{h} W_h \, \bar{y}_{lrh} \qquad (2)$$

$\bar{y}_{lrs}$ is an unbiased estimate of the population mean ($\bar{Y}$) .

This estimate is appropriate when it seems to us that the regression coefficient for the community, Bh, changes from stratum to stratum.

Since the sampling is independent in the different strata, the variance of this estimate is:

$$V\,(\bar{y}_{lrs}) = \sum_{h} \frac{W_h^2\,(1 - f_h)}{n_h} \left(\sigma_{yh}^2 - 2b_h\,\sigma_{yxh} + b_h^2\,\sigma_{xh}^2\right) \qquad (3)$$

It is the smallest possible when it is equal to the regression coefficient of the population, and thus the smallest variance is as follows**:**

$$V\,(\bar{y}_{Lrs}) = \sum_{h} \frac{W_h^2\,(1 - f_h)}{n_h} \left(\sigma_{yh}^2 - \frac{\sigma_{yxh}^2}{\sigma_{xh}^2}\right) \qquad (4)$$

The first degree of approximation to the MSE ($\hat{\bar{Y}}_{lrs}$) is as follows :

$$\mathrm{MSE}\left(\hat{\bar{Y}}_{Lrs}\right) \cong \sum_{h=1}^{L} W_h^2\,\lambda_h\left[\sigma_{yh}^2 + \beta_h^2\,\sigma_{xh}^2 - 2\beta_h\sigma_{yxh}\right] \qquad (5)$$

Where:

Sampling break within stratum *h*.

$$f_h = \frac{n_h}{N_h}$$

σ_{yh}^2 is the population variance of the study variable in stratum h .

$$\boldsymbol{\sigma_{yh}^2 = \frac{\sum_{i=1}^{N_h}(y_{hi} - \bar{Y}_h)^2}{N_h - 1}}$$

σ_{xh}^2 is the population variance of the auxiliary variable in class h.

$$\sigma^2_{xh} = \frac{\sum_{i=1}^{N_h}(x_{hi} - \overline{X}_h)^2}{N_h - 1}$$

σ_{yxh} is the covariance between the covariate and the study variable in class h.

$$\sigma_{xyh} = \frac{\sum_{i=1}^{N_h}(x_{hi} - \overline{X}_h)(y_{hi} - \overline{Y}_h)}{N_h - 1}$$

4. Outliers in data

The challenge of outliers has long been acknowledged as a critical issue in statistics. It presents a significant obstacle for regression estimation techniques, as the inclusion of outliers in a dataset can profoundly affect the statistical analysis and subsequent decision-making processes. Outliers are observations that markedly deviate from the majority of the data and are also referred to as extreme values, discordant observations, rogue values, contaminants, mavericks, or dirty values. Awareness of the outlier problem dates back to the early days of the ordinary least squares (OLS) method. (Begashaw, and Yohannes. 2020)

5. Robust Regression

Robust regression comprises a variety of techniques employed to construct regression models that are resilient to violations of the assumptions in classical linear regression. Such violations may involve the presence of outliers or heavy-tailed distributions, among other factors.

The main goal of robust regression methods is to derive regression coefficients that are less influenced by aberrant observations. This approach ensures more reliable conclusions and predictions. Robust regression techniques often utilize alternative measures of fit and estimation, such as the Hampel M, Tukey M, and Least Trimmed Squares (LTS) estimators. These estimators are designed to minimize the effect of outliers, thereby enhancing the robustness of the regression analysis.

Robust regression methods become especially valuable when confronted with datasets characterized by high variability or containing outliers that aren't merely attributable to measurement errors or missing data, but instead signify genuine deviations from the assumed model.(Maronna,R.A.,et al,2006)

6. Breaking point

The "breaking point" refers to the minimum contamination percentage or fraction of remote observations at which an anomaly detection method begins to recognize outliers. Essentially, it represents the threshold beyond which the method's effectiveness diminishes due to an excessive number of outliers or contaminated data points. Understanding the breaking point is critical in robust statistics as it quantifies the method's ability to manage outliers without sacrificing reliability. (Maronna, R.A., et al, 2006).

7. Robust Estimators

Robust methods for discrete regression parameter estimation involve techniques that aim to estimate regression parameters while minimizing the influence of outliers or violations of model assumptions. These methods are particularly useful when traditional regression techniques, such as ordinary least squares (OLS), are sensitive to deviations from the assumed model structure. Some common robust methods for discrete regression parameter estimation include:

1. Hampel's M-estimator, introduced in 1971, is a robust statistical method used for parameter estimation. It minimizes a robust criterion function that reduces the impact of outliers on the estimation process.

$$\mathrm{u}_{Hampel}(\mathrm{u},\mathrm{c}) = \begin{cases} 1 & if\ |u| \le a \\ \dfrac{a}{|u|} & if\ a < |u| \le b \\ \dfrac{a(c-|u|)}{|u|\,(c-b)} & b < |u| \le c \\ 0 & otherwise\ (i.e. |u| > c \end{cases} \qquad (6)$$

Where (c, a , b) is a constant.

2. Tukey's bisquare M-estimator: Similar to the Huber M-estimator, this method minimizes a combination of squared and absolute errors. However, it assigns smaller weights to outliers than the Huber M-estimator, making it more robust to extreme deviations from the model.

$$u_{Tukey} = \begin{cases} u(1-u^2)^2 & ; \quad |u| < c \\ 0 & ; \quad |u| \geq c \end{cases} \qquad (7)$$

Where C is a constant that takes values (4.685 , 6.0). (TUKEY, J.W. 1983)

3. Least Trimmed Squares (LTS): LTS involves fitting separate regression lines to subsets of the data, removing a specified percentage of observations with the largest residuals. This method is particularly robust to outliers because it focuses on the observations that contribute the least to the overall fit. (Zaman and Bulut. 2018)

$$\min \sum_{i=1}^{z} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2 \qquad (8)$$

where: $\boldsymbol{z = \frac{n}{2+1}}$

4. The Minimum Volume Ellipsoid Estimator (MVE) is a robust statistical method used for estimating parameters in multivariate data sets, particularly in the presence of outliers. It is a geometric approach that seeks to fit an ellipsoid of minimum volume around the central core of the data, effectively capturing the majority of the observations while minimizing the influence of outliers. (Peter Rousseeuw , 1985)
The MVE method aims to find the ellipsoid that contains a specified proportion of the data points with the smallest volume. This ellipsoid is centered at the centroid of the data and is determined by solving an optimization problem that minimizes the volume subject to certain constraints. These constraints typically ensure that the ellipsoid encloses a predetermined proportion of the data points and has a specified shape, such as being axis-aligned or having a certain aspect ratio.

The MVE estimator is robust to outliers because it focuses on the core of the data rather than the entire data set. Outliers have a limited impact on the estimation process because they do not significantly affect the shape and size of the minimum volume ellipsoid.

One advantage of the MVE estimator is its ability to handle high-dimensional data sets where traditional methods may struggle due to the curse of dimensionality. However, the computational complexity of finding the minimum volume ellipsoid increases with the dimensionality of the data, which can be a drawback for very high-dimensional data sets. (Rousseeuw , 1985 ; Zaman and Bulut. 2023)

5. The Minimum Covariance Determinant (MCD) estimator is a robust statistical method used for estimating the covariance matrix of multivariate data sets, particularly in the presence of outliers. It is based on the concept of robust estimation of location and scatter, where the goal is to find the center and spread of the data that are not unduly influenced by outliers. (Rousseeuw , 1985)

   The MCD estimator seeks to identify a subset of data points (referred to as the Minimum Covariance Determinant subset) that have the minimum determinant of the covariance matrix among all possible subsets of a specified size. Intuitively, this subset represents the most "normal" or "well-behaved" portion of the data, and its covariance matrix provides a robust estimate of the underlying covariance structure.

   Once the Minimum Covariance Determinant subset is identified, the MCD estimator computes the covariance matrix using these data points. This covariance matrix is less affected by outliers compared to traditional estimators, such as the sample covariance matrix or the maximum likelihood estimator, which are sensitive to the presence of outliers. (Rousseeuw, 1987)

   The MCD estimator is particularly useful in applications where the data may contain outliers or heavy-tailed distributions, such as finance, environmental monitoring, and quality control. It provides a robust and efficient way to estimate the covariance matrix, which is a fundamental parameter in many statistical analyses, including multivariate regression, principal component analysis, and discriminant analysis.(Rousseeuw , 1987)

   One limitation of the MCD estimator is that it can be computationally intensive, especially for large data sets or high-dimensional data. However, various algorithms and implementations have been developed to address this issue and make the MCD estimator practical for real-world applications. Overall, the Minimum Covariance Determinant estimator is a valuable tool for robust covariance estimation in multivariate data analysis. (

Rousseeuw (1985) ; Rousseeuw and van Driessen 1999 ; Bulut and Oner 2017 ; Bulut 2014)

8. Criteria for comparing robust methods

In the pursuit of identifying the most effective estimator, various criteria (scales) have been established for comparing estimation methods, with preference given to those demonstrating the least error. Consequently, numerous comparison scales are employed among immune estimators, one of which is:

$$\mathrm{MSE}\left(\hat{\hat{\mathrm{Y}}}_{\mathrm{Lrs}}\right) \cong \sum_{h=1}^{L} W_h^2\, \lambda_h \left[\sigma_{yh}^2 + \beta_h^2\, \sigma_{xh}^2 - 2\beta_h \sigma_{yxh}\right] \qquad (9)$$

$$\mathrm{MSE}\left(\hat{\hat{\mathrm{Y}}}_{\mathrm{Lrs(i)}}\right) \cong \sum_{\mathrm{h}=1}^{\mathrm{L}} \mathrm{W}_{\mathrm{h}}^2\, \lambda_{\mathrm{h}} \left[\sigma_{\mathrm{yh}}^2 + \beta_{\mathrm{hrobust(s)}}^2\, \sigma_{\mathrm{xh}}^2 - 2\beta_{\mathrm{hrobust(s)}}\, \sigma_{\mathrm{yxh}}\right] \qquad (10)$$

$$\mathrm{MSE}\left(\hat{\hat{\mathrm{Y}}}_{\mathrm{Lrs(ig)}}\right) \cong \sum_{\mathrm{h}=1}^{\mathrm{L}} \mathrm{W}_{\mathrm{h}}^2\, \lambda_{\mathrm{h}} \left[\sigma_{\mathrm{yh(g)}}^2 + \beta_{\mathrm{hrobust(z)}}^2\, \sigma_{\mathrm{xh(g)}}^2 - 2\beta_{\mathrm{hrobust(z)}}\, \sigma_{\mathrm{yx(g)}}\right] \qquad (11)$$

A comparison can be made between the two standards MSE for $\hat{\hat{\mathrm{Y}}}_{\mathrm{Lrs}}$ and MSE for ($\hat{\hat{\mathrm{Y}}}_{\mathrm{Lrs\,(i)}}$ as follows :

$$\mathrm{MSE}\left(\hat{\hat{\mathrm{Y}}}_{\mathrm{Lrs(i)}}\right) < \mathrm{MSE}\left(\hat{\hat{\mathrm{Y}}}_{\mathrm{Lrs}}\right) \qquad (12)$$

When the condition is met

(12): $\left(\tau_{(g)} - \tau\right) - 2\left(\delta_{(g)} - \delta\right) < 0$

Whereas

$$\tau_{(g)} = \sum_{h=1}^{L} \beta_{hrobust(s)}^{2} W_h^2 \lambda_h \sigma_{xh(g)}^{2}$$

$$\delta_{(g)} = \sum_{h=1}^{L} \beta_{hrobust(s)} W_h^2 \lambda_h \sigma_{yxh(g)} \quad ; \quad (\text{g} = \text{MCD and MVE}) ; \ (\text{s} = \text{Hampel M, Tukey M, LTS})$$

$$\tau = \sum_{h=1}^{L} \beta_h^2 W_h^2 \lambda_h \sigma_{xh}^2 \quad , \quad \delta = \sum_{h=1}^{L} \beta_h W_h^2 \lambda_h \sigma_{yxh}$$

$$\tau_{(r)} = \sum_{h=1}^{L} \beta_{hrobust(s)}^{2} W_h^2 \lambda_h \sigma_{xh}^{2} \quad , \quad \delta_{(r)} = \sum_{h=1}^{L} \beta_{hrobust(s)} W_h^2 \lambda_h \sigma_{yxh}$$

The estimates of robust regression methods for discrete regression (Hampel M, Tukey M, and Least Trimmed Squares (LTS)) in which the regression parameter β was employed to estimate the population mean in stratified random sampling, using the traditional covariance and covariance matrix in the MSE formula (10) where it is more efficiency.

Likewise if :

$$\text{MSE}\left(\hat{\hat{\bar{Y}}}_{\text{Lrs(sg)}}\right) < \text{MSE}\left(\hat{\hat{\bar{Y}}}_{\text{Lrs}}\right)$$

When the condition is fulfilled (13)

$$\left(\text{T}_{(\text{g})} - \text{T} + \tau_{(g)} - \tau\right) - 2\left(\delta_{(g)} - \delta\right) < 0 \qquad (13)$$

The estimators of the robust regression methods of discrete regression (Hampel M, Tukey M, and Least Trimmed Squares (LTS)) in which the regression parameter β was employed to estimate the population mean in the stratified random sampling, using the robust covariance and covariance matrix of the estimators (MCD, MVE) in the MSE formula (11) more efficient.

9. Applied Side

In this context, we utilized the R programming language (version 4.1.3) for statistical analysis to meet the research objectives of comparing robust estimators and identifying the most suitable ones based on efficiency and mean squared error criteria.

For the purpose of applying the aim of this research in verifying the performance of robust estimation methods, we have studied real data from a comprehensive survey of the number of hotels in Dohuk Governorate, where their data were taken through the Central Bureau of Statistics and included the following information:

X : Number of employees .

Y : Wages paid in thousands of dinars .

9.1 Results of applying the robust methods :

Table .1 shows descriptive statistics for the hotel population

Population		Strata	1	2	3	4	5
N	115	N_h	26	13	8	24	44
N	66	n_h	15	7	5	14	25
$\bar{X}$	4.904348	$\bar{X}_h$	10.19231	5.153846	4.25	3.958333	2.340909
$\bar{Y}$	13084.61	$\bar{Y}_h$	8970.696	1111.304	490.4348	1321.739	1190.435
S_x	6.527908	S_{xh}	11.92315	3.131724	2.37547	1.398109	1.18013
S_y	38564.51	S_{yh}	75799.19	6827.931	4602.717	4048.34	3523.012

r_{xy}	0.9440883	r_{xyh}	0.9570255	0.9467657	0.4906233	0.7930014	0.7385142
$r_{xy.out}$	0.8158777	$r_{xyh.ou}$	0.8782181	-0.1471759	0.0363727 6	0.3150866	0.7385142
$\bar{Y}_o$	21903.83	$\bar{Y}_{h.out}$	53243.85	25961.54	35062.5	15820.83	3111.364
S_y	45393.26	$S_{yh.out}$	79284.53	26834.91	28231.69	29068.13	3523.012
		λ_h	0.02820513	0.06593407	0.075	0.0297619	0.01727273
		w_h	0.226087	0.1130435	0.06956522	0.2086957	0.3826087

Table .2 displays the (RE, MSE) values for estimating the population mean using discrete regression with robust regression methods, employing both traditional covariance and covariance matrix approaches

Methods	Wages paid in thousands of dinars (Y)	
	MSE	RE
OLS	7065346	1
LTS	8014238	0.88159922
Tukey M	7550898	0.93569613
Hampel M	7538197	0.93727267

From the data presented in the above table, it is evident that the robust regression methods (LTS, Tukey M, Hampel M) utilized for estimating the regression parameter β in stratified discrete regression, using both traditional covariance and covariance matrices according to formula (10) for population mean estimation, do not exhibit higher efficiency compared to the traditional least squares method (OLS). This is indicated by the larger values of MSE associated with the robust estimators, rendering them less efficient. This outcome was anticipated given that the conditions outlined in formula (12) are not satisfied for all the aforementioned estimators in this dataset.

This discrepancy becomes apparent upon reviewing Table (3), where it becomes evident that robust estimators should outperform traditional estimators in terms of efficiency if the conditions specified in formula (12) are met.

Table .3 presents the outcomes of the condition specified in formula (12)

Methods	Wages paid in thousands of dinars (Y)	
	Condition values	Results
LTS	9.49E+05	FALSE
Tukey M	4.86E+05	FALSE
Hampel M	4.73E+05	FALSE

Table 4 displays the (RE, MSE) values for estimating the population mean using discrete regression with robust regression methods, utilizing robust covariance and covariance matrix approaches, namely MCD and MVE

Methods	Wages paid in thousands of dinars (Y)			
	MSE		RE	
	MCD	MVE	MCD	MVE

OLS	453302.4	459043.5	1	1
LTS	109113.4	116986.7	4.1544155	3.923895
Tukey M	126025.7	133855.4	3.59690444	3.429398
Hampel M	129347.8	137686.0	3.50452346	3.333988

In Table .4, in accordance with formula (11), it is observed that the efficiency criterion for estimates obtained through robust regression methods (LTS, Tukey M, Hampel M), utilized to estimate the regression parameter β using robust covariance and covariance matrices for calculating (s_{xy} , s_y^2 , s_x^2) via estimators (MCD, MVE) to estimate the population mean via the stratified discrete regression method

($\hat{\hat{Y}}_{\mathrm{lrs}}$), exceeds 1. This suggests that estimates derived from robust methods exhibit greater efficiency compared to the traditional least squares (OLS) method.

Furthermore, the results in the table indicate that the estimation achieved by LTS utilizing the hippocampus's variance-covariance matrix (MCD) is more efficient than the other methods, as evidenced by its lowest MSE value. This outcome is anticipated since the conditions outlined in formula (13) are satisfied for all estimates of robust regression methods (LTS, Tukey M, Hampel M) within the dataset, as clearly illustrated in Table .5.

Table .5 presents the outcomes of the condition specified in formula (13)

Methods	Wages value (Y) in thousands dinars			
	MCD	Results	MVE	Results
LTS	- 6956232.6	TRUE	- 6948359.3	TRUE
Tukey M	- 6939320.3	TRUE	- 6931490.6	TRUE

Hampel M	- 6935998.2	TRUE	- 6927660	TRUE

10. Conclusions

Utilizing robust regression estimates and robust covariance matrices in stratified random sampling improves the efficiency of estimates for the population mean compared to traditional methods when dealing with outliers in datasets. This efficiency advantage holds true across various proportions of outliers and sample sizes. Robust estimators consistently outperform traditional discrete regression estimators in terms of efficienc.

References

1. William G. Cochran , (1977) , " Sampling Techniques " , 3rd. edition , John Wiley & Sons , Inc. ,USA .

2. Zaman, T., and H. Bulut. 2018. " Modified ratio estimators using robust regression Methods " , Communications in Statistics - Theory and Methods .

3. Bulut, H., and Y. Oner. 2017." The evaluation of socio-economic development of development agency regions in Turkey using classical and robust principal component analyses". Journal of Applied Statistics 44 (16):2936–48.

4. Rousseeuw, P.J., Leroy ,A. M.(1987) " Robust Regression and Outliers Detection " John Wiley , New York .

5. Rousseeuw, P. J., Driessen, K. V. (1999) "A fast algorithm for the minimum covariance determinant estimator". Technometrics, vol 41(3) , pp. 212-223

6. Dick J. Brus . " Spatial Sampling with R " . CRC Press is an imprint of Taylor & Francis Group, LLC, Boca Raton London New-York , 2022 .

7. Begashaw, and Yohannes. (2020): " Review of Outlier Detection and Identifying Using Robust Regression Model " . International Journal of Systems Science and Applied Mathematics ; 5(1) : 4 -11

8. Maronna, R. A., Martin, R. D. and Yohai, V. J. (2006), " Robust Statistics ". John Wiley .

9. Ahmed, Rikan A., and Saja Mohammad Hussein. "Some ratio estimators of finite population variance using auxiliary information in ranked set sampling." *International Journal of Nonlinear Analysis and Applications* 13.1 (2022): 3537-3549. Baghdad

10. Zaman, T., and H. Bulut. 2023. " Robust calibration for estimating the population mean using stratified random sampling " , Communications in Statistics - Theory and Methods .

11. Bulut, H. 2014." Use of robust statistics in multivariate statistical analysis". Master thesis, Ondokuz Mayıs University.

12. HOAGLEN, D.C., MOSTELLER, F & TUKEY, J.W. (1983) Understanding Robust and Exploratory Data Analysis. Wiley, New York.