Journal of Economics and Administrative Sciences (JEAS)

Available online at http://jeasiq.uobaghdad.edu.iq

Estimating the Population Mean in Stratified Random Sampling Using Combined Regression with the Presence of Outliers

Mustafa Habib Mahdi
Department of Statistics /College of Administration and Economics / University of Baghdad
Baghdad, Iraq
mostafa.Habeeb1201a@coadec.uobaghdad.edu.iq

Saja Mohammad Hussein
Department of Statistics /College of Administration and Economics / University of Baghdad
Baghdad, Iraq
Saja@coadec.uobaghdad.edu.iq

Received: 28/3/2023 **Accepted: 8/5/2023** **Published: June / 2023**

Abstract

In this research, the covariance estimates were used to estimate the population mean in the stratified random sampling, and combined regression estimates. were compared by employing the robust variance-covariance matrices estimates with combined regression estimates by employing the traditional variance-covariance matrices estimates when estimating the regression parameter, through the two efficiency criteria (RE) and mean squared error (MSE). We found that robust estimates significantly improved the quality of combined regression estimates by reducing the effect of outliers using robust covariance and covariance matrices estimates (MCD, MVE) when estimating the regression parameter. In addition, the results of the simulation study proved that the Minimum covariance determinant (MCD) method is highly efficient at all sample sizes (n=35 , 75 , 150 , 200 , 500) and then followed by the method of the smallest ellipse Minimum volume Ellipsoid (MVE) handles outliers in the dataset, where it has lower values (MSE).

Paper type: Research paper.
Keywords: Combined regression estimator, outliers, Minimum covariance determinant (MCD) and Minimum volume Ellipsoid (MVE) , relative efficiency, stratified random sampling.

1. Introduction

Stratified random sampling (STRS) is a sampling in which the population units are completely heterogeneous, then some restrictions can be placed on SRS to increase the accuracy of the estimate, by reducing the effect of heterogeneity.

The simplest constraint is the division of society into partial groups called strata, as each stratum is homogeneous within it and differs from the other strata. And when determining the number of layers, a sample is drawn from each layer, and the drawing is done independently and randomly in the different layers. (Al-Nasser & Al-Safawi, 2001)

In stratified random sampling, estimating the population mean for a variable of interest is often important. This can be done using linear regression, which can estimate the mean using the information from the strata. However, the standard linear regression estimator assumes that the regression coefficients are the same across all strata, which may not be accurate in practice. In such cases, alternative linear regression estimators can be used, which consider the heterogeneity of regression coefficients across strata. (Rikan & Saja, 2022)

One of these estimators is the combined regression estimator. This estimator is used in regression analysis to estimate the relationship between the variable of interest and the stratification variable, and then combines the estimates from each stratum to create an overall estimate.

However, one of the problems that can arise with the combined regression estimator is the presence of outliers, outliers are data points that differ significantly from the rest of the data, and can have a significant impact on the regression analysis.

In the context of stratified random sampling, outliers can occur within each stratum or in the overall sample. An outlier in one stratum can skew the regression analysis and lead to an inaccurate estimate of the community coefficient for that stratum. Similarly, suppose an outlier is present in the total sample. In that case, it can have a significant impact on the pooled regression estimator and lead to an inaccurate estimate of the population coefficient.

There are many studies on the characteristics of the combined regression estimator in stratified random sampling, including:

Gupta and Singh (1983) studied the properties of the combined regression estimator in stratified random sampling under simple stratified random sampling, where they derived the variance of the combined regression estimator and compared it with the variances of the usual estimators. They concluded that the combined regression estimator is generally more efficient than other estimators, especially when the correlations between Strata are high.

Kim and Lee (2005) studied the joint regression estimator in stratified random sampling with unequal sample sizes. They derived the variance of the combined regression estimator and compared it with the variances of other estimators. They found that the combined regression estimator is more efficient than the other estimators when the sample sizes are unequal.

Alwan et al (2011) discussed the comparison between the variance of the simple random sampling average and the variance of the stratified random sampling average represented by the variance of the average of the separate regression estimator and the variance of the average of the joint proportion estimator$cumf^{\frac{4}{5}}$.

Chaudhary and Singh (2014) studied the properties of the co-regression estimator in stratified random sampling under unequal probability sampling of strata. They derived the variance of the combined regression estimator and compared it with the variances of other estimators. They concluded that the combined regression estimator is more efficient than the other estimators, especially when the sample sizes are small.

Shrestha and Singh (2016) studied the properties of the combined regression estimator in a stratified random sampling with no response. They derived the variance of the co-regression estimator and compared it with the variances of other estimators. They found

that the combined regression estimator is more efficient than other estimators, especially when nonresponse rates are high.

The two researchers Rikan and Saja (2022) presented a proposal aimed at estimating the average number of the limited community for the main variable by means of the stratification group of the STRSS sample through the modification that was made to the exponential estimator for the type of documents and the generalized product.

The relative bias PRB, mean squared error MSE and percentage relative efficiency PRE of the proposed modified estimator were obtained to the first degree of approximation.

The objective is to estimate the population mean by covariance method with the use of the robust covariance and covariance matrix using(Minimum covariance determinant (MCD) estimator, Minimum volume ellipsoid (MVE) estimator) to estimate the regression parameter, in stratified random sampling in the presence of outliers in the data set.

Comparing the covariance estimates by employing the robust covariance and covariance matrix estimates with the covariance estimates by employing the traditional covariance matrix estimates when estimating the regression parameter, through the two criteria of efficiency (RE) and mean square error (MSE).

Regression analysis is a formidable tool for exploring and understanding the relationship between variables; however, this analysis is sensitive to outliers, which are data points that deviate significantly from the overall trend, outliers can have a significant effect on the estimated regression coefficient and its dependence when estimating a parameter on the covariance and variance matrix Subscriber.

When outliers are present in the data, they can skew the estimated regression line and cause the slope of the line to be inaccurate. This can lead to misleading predictions and conclusions about the relationship between the variables. Outliers can also increase the variance of the estimated regression coefficient, causing It makes it less accurate and reliable.

In addition, the effect of outliers on the regression coefficient depends on the position of the anomaly relative to other data points. If the outlier is close to other data points, it may have little effect on parameter estimation. However, if the outlier is far from other data points, it may greatly affect parameter estimation and lead to erroneous conclusions.

The effect of outliers on the covariance matrix and covariance can increase the covariance between variables, making it difficult to determine the true relationship between variables. Also, outliers can inflate the covariance matrix, which may lead to inaccurate and unreliable regression coefficient estimates. And associated standard errors.

When the data contains outliers, the traditional estimators can be affected and their efficiency can be reduced, which is why it has become necessary to use the robust covariance and covariance matrix when working with polluted or heterogeneous data, where data points in a given group do not follow the same distribution. Supposed, in such cases, several estimates are available for this type of data, namely the Minimum covariance determinant (MCD) and Minimum volume ellipsoid (MVE) estimators.

2. Materials and Methods

2.1 Estimators of Linear regression in Stratified Random Sampling

One of the methods for estimating the population mean in stratified random sampling is based on the linear regression of Yi over Xi, where one of the types of these estimators is: (Rousseeuw, P.J., 1985)

2.1.1 Combined regression Estimator ($\bar{y}_{lrc}$)

A combined regression estimator is a statistical method used in stratified random sampling, which allows more accurate estimates of population parameters by integrating information from different regression models. (William G. Cochran, 1977)

The combined regression estimator is a modification of the standard regression estimator, in which the regression coefficients are estimated separately within each stratum. Then, the coefficients estimated across the stratua are combined to obtain an overall estimate of the population coefficient. This approach allows a better estimation of population parameters, especially when significant differences between strata in terms of variables of interest exist.

In summary, the combined regression estimator is a useful tool in stratified random sampling that can improve the accuracy of estimating population parameters by considering the variance between strata. In addition, it is beneficial when there are strong correlations between the stratification variables and the variables of interest, which can lead to biased estimates if ignored.

The formula for estimating the mean using the combined regression method is:

$$\hat{\bar{Y}}_{lrc} = \bar{y}_{st} + b_c\,(\bar{X} - \bar{x}_{st}) \qquad \dots (1)$$

Where as

$$\bar{y}_{st} = \sum_{h=1}^{L} W_h \bar{y}_h \quad , \quad \bar{x}_{st} = \sum_{h=1}^{L} W_h \bar{x}_h \quad , \quad L = 1,2,3,4,5$$

$$\bar{y}_h = \frac{\sum_{i=1}^{n_h} y_{hi}}{n_h} \quad , \quad \bar{x}_h = \frac{\sum_{i=1}^{n_h} x_{hi}}{n_h} \quad , \quad W_h = \frac{N_h}{N}$$

$$\bar{X} = \frac{\sum_{h=1}^{L}\sum_{i=1}^{N_h} x_{hi}}{N} = \frac{\sum_{h=1}^{L} N_h\ \bar{X}_h}{N} = \sum_{h=1}^{L} W_h\ \bar{X}_h$$

We can see from formula (1) that $\hat{\bar{Y}}_{lrc}$ is also an unbiased estimate of the population mean $\bar{Y}$.

Since $\hat{\bar{Y}}_{lrc}$ is an estimator from a stratified sample of the variable $y_{hi} + b\,(\bar{X} - x_{hi})$ the variance of the estimator is as follows :

$$V\left(\hat{\bar{Y}}_{lrc}\right) = \sum_{h} \frac{W_h^2\,(1 - f_h)}{n_h}\,\left(S_{yh}^2 - 2b_h\,S_{yxh} + b^2\,S_{xh}^2\right) \qquad \dots (2)$$

The value of b that makes the variance as small as possible is:

$$b_c = \frac{\sum_h \frac{W_h^2\,(1 - f_h)\,s_{yxh}}{n_h}}{\sum_h \frac{W_h^2\,(1 - f_h)\,s_{xh}^2}{n_h}} \qquad \dots (3)$$

The first order of approximation to the MSE ($\hat{\bar{Y}}_{lrc}$) is given by : (Willis A . Jensen et al., 1987)

$$\mathrm{MSE}\left(\hat{\bar{Y}}_{lrc}\right) \cong \left\{\sum_{h=1}^{L} W_h^2 \lambda_h\, S_{yh}^2 + \beta_c^2 \sum_{h=1}^{L} W_h^2 \lambda_h S_{xh}^2 - 2\beta_c \sum_{h=1}^{L} W_h^2\,\lambda_h\, S_{yxh}\right\} \quad (4)$$

Where

$\beta_c = \frac{\sum_{h=1}^{L} W_h^2\,\lambda_h\,S_{yxh}}{\sum_{h=1}^{L} W_h^2\,\lambda_h\,S_{xh}^2}$ is obtained by the classic covariance matrix.

$\lambda_h = \frac{1 - \frac{n_h}{N_h}}{n_h}$

nh is the number of units in sample stratum h.

S_{yh}^2 is the population variance of variate of interest in stratum h.

S_{xh}^2 is the population variance of auxiliary variate in stratum h.
S_{yxh} is the population covariance between auxiliary variate and variate of interest in stratum h.
And these parameters are computed by using a classic covariance matrix.
It is obtained through the classic covariance matrix.

2.2 Outliers

The outliers problem is one of the most important and oldest problems in statistics. Perhaps it is one of the main problems of regression estimation methods, where the appearance of outliers in the data set affects the result of the statistical analysis of the data, that the outliers are the observations that are very different from the other observations that are supposed to be the result on a different mechanism. (Dick, 2022)

The effect of anomalous observations within a single sample on the properties of the ordinary least squares estimators, as shown by drawing how these observations change the trend line of least squares. (William G. Cochran, 1977)

2.3 Robust Methods for Estimating the Covariance Matrix and Covariance of the Combined Regression Parameter. (Hubert, 2004)

2.3.1 Minimum Volume Ellipsoid Estimator (MVE)

In robust regression analysis, the smallest ellipse (MVE) estimator is often used to estimate the location and measurement parameters in multiple variables. For the variance and covariance matrix, this estimator is defined as the smallest ellipse containing half of the observations, while the location estimate in (MVE) is the midpoint. So the ellipse, and this estimator have a breakdown point is $\frac{(\left|\frac{n}{2}\right|-P+1)}{n}$. And when it is not n → ∞, the breakdown point of this estimator (MVE) is 50%. (Rousseeuw, 1985)

The algorithm for the MVE estimator can be followed by the following:

1. Several of subsamples equal to C_{p+1}^{n} must be selected for the original data matrix containing n observations (rows) and p variables (columns).

2. Calculate the value of each of the arithmetic means ($\mu_{(j)}$) and the variance matrix ($\Sigma_{(j)}$) for each partial sample (j) with size (p+1).

3. Calculate the square distances for each of ($\Sigma_{(j)}$, $\mu_{(j)}$) according to the following formula:

$$D_i^2(j) = (x_i - \mu_j)' \Sigma_j^{-1} (x_i - \mu_j)$$

4. Arrange the square distances according to the third step in ascending order, and then determine the value of $D_{hp}^2(j)$ as:

$$hp = \frac{n+p+1}{2}$$

For each partial sample (j).

5. Finding the best selected partial sample (j), with the smallest objective function.

$$\tilde{j} = \underset{j}{arg\ min\ det} \sum_{j} D_{hp(j)}^2$$

And then determine the mean **μ** and variance **Σ** for this partial sample.

6. Find the measurement matrix correction coefficient (dispersion) by:

$$C = 1 + \frac{15}{n-p}$$

7. Find the amplified variance matrix according to the following formula:

$$S(x) = (\chi_{p,0.5}^2)^{-1} * \sum_{j} D_{hj}^2 \qquad (5)$$

8. Finding the robust distances again according to Equation 4 and for all observations of the sample (n), and according to the following formula:

$$d^2_{i(j)} = (x_i - \mu_{(j)})' S^{-1}(x)(x_i - \mu_{(j)}) \quad (6)$$

9. Adding one more observation to the subgroup to be (p+2) is often the same observations for the first group (p+1) with the addition of a new observation.

10. From step (9), by choosing the partial sample (p+2), steps from (2) to step (9) are repeated to obtain a sample of size (p+2) from the values.

11. Continue to increase each observation to each stage until the stopping limit is reached, which is the size of the partial sample (hp). This sample is homogeneous and consistent with each other, and its direction is towards the center.

Knowing that the breakdown point of the MVE estimator is equal to **h=(1-**α **) *n**, where (α) is the contamination percentage of the studied data.

12. Calculating the arithmetic mean vector $\mu_{(hp)}$ and the measurement matrix (dispersion) $\Sigma_{(hp)}$ for the sample calculated from the eleventh step (hp) by finding the squared fortified distances and diagnosing the outliers through it (which is the value that increases the value of the fortified distance for the value of ($\chi^2_{p,0.975}$).

2.3.2 Minimum Covariance Determinant Estimator (MCD)

A robust statistical method was developed by Peter Rousseeuw in 1985, the MCD estimator is used to estimate the center and variance of a multivariate data set in the presence of outliers. (Hawkins, 1980)

The main idea behind the MCD estimator is to find a subset of the data, called a subset of the minimum covariance, with the smallest possible determinant of variance. This subset estimates the center and variance of the entire data set. (Dick, 2022)

The MCD estimator is particularly useful when the data may contain outliers, as it is impervious to up to half of the observations being outliers. In addition, the MCD estimator can be used with high-dimensional data, where the number of variables is much greater than the number of observations. (Rousseeuw, 1985)

The algorithm for finding MCD estimators for both location and dispersion is as follows:

1. Select all possible subsamples **J** of size **h** from the observations using the formula:

$$C^n_h = \frac{n!}{h!\,(n-h)!} \qquad J = i_1, i_2, \dots\dots, i_h \qquad (7)$$

Whereas

n: is the total number of observations.

h: represents approximately half of the data and is equal to**:**

$$\mathbf{h} = \frac{n + k + 1}{2} \qquad (8)$$

2. Calculate the arithmetic mean vector **($\underline{\hat{\mu}}$)** and covariance matrix **(cv)** for each subsample **J** using the following formulas:

$$\underline{\hat{\mu}} = \frac{1}{h}\sum_{i=1}^{h} x_j \qquad (9)$$

$$\mathbf{cv} = \frac{1}{h-1}\sum_{i=1}^{h}\left(x_j - \underline{\hat{\mu}}\right)\left(x_j - \underline{\hat{\mu}}\right)' \qquad (10)$$

3. Finding the best sub-sample J, which is the sample that has the smallest determinant of the covariance matrix among all samples by**:**

$$\mathcal{J} = arg_J\ min\,|cv| \qquad (11)$$

Multiply the estimates of the arithmetic mean vector and covariance matrix for the best subsample by the consistency factor **($C_{n,k}{}^2$)** used in the MVE method to obtain the estimates for the smallest determinant of the covariance matrix (MCD), from Then calculate the square distance (Mahalanobis) for all observations in the sample using the estimated values through the following formula: (Rousseeuw, 2018).

$$\boldsymbol{RMD_{iMCD}}^2 = \left(\boldsymbol{x_i} - \underline{\hat{\boldsymbol{\mu}}}_{\boldsymbol{MCD}}\right) (\boldsymbol{cv_{MCD}})^{-1} \left(\boldsymbol{x_i} - \underline{\hat{\boldsymbol{\mu}}}_{\boldsymbol{MCD}}\right)' \quad (\mathbf{12})$$

2.4 Comparison Criteria

To reach the most efficient estimator, several criteria were found for comparison between the estimation methods, and the best one is the one with the least possible error. One of these criteria is the mean square error (MSE). (Rousseeuw, 1987)

$$MSE = \frac{1}{R} \sum_{i=1}^{R} (\hat{\bar{Y}}_i - \bar{Y})^2 \qquad i = 1,2,\dots,k \qquad \dots. \quad (13)$$

R: represents the number of iterations for each experiment (500) samples.

$\hat{\bar{Y}}_i$: The estimated value of the population mean of the dependent variable.

$\bar{Y}$: The true value of the population mean of the dependent variable.

$$RE\left(\hat{\bar{Y}}_{lrc(q)}\right) = \frac{MSE(\hat{\bar{Y}}_{lrc})}{MSE(\hat{\bar{Y}}_{lrc(q)})} \qquad (14)$$

Whereas q = MCD and MVE

3. Discussion of Results

3.1 Simulation Preparation

In this aspect, we used the R language program (version (4.1.3)) in conducting the statistical analysis to achieve the study's objectives. To compare robust estimators and test the best ones based on the standard of mean squares error.

The simulation experiments are described as follows:

1.The default values of the combined regression parameter were determined from the real data by estimating them using the (OLS) method.

2.Selection of five different sizes of assumed samples (35, 75,150, 200,500) from a total population of size (N=700).

And one explanatory variable was determined for each of the (5) strata, where the simple regression model used for each stratum was:

$$Y_1 = 3 + 4X_1 + \varepsilon$$
$$Y_2 = 7 + 3X_2 + \varepsilon$$
$$Y_3 = 12 + X_3 + \varepsilon$$
$$Y_4 = 18 + 8X_4 + \varepsilon$$
$$Y_5 = 5 + 7X_5 + \varepsilon$$

3.Generate explanatory variables X_{ij} for five explanatory layers (variables) and adopt different percentages of pollution (10%, 25%, 30%, 35%, 5%).

$$X_1 \sim N(2,2)\ , X_2 \sim N(0,4), X_3 \sim N(7,3), X_4 \sim N(3,11), X_5 \sim N(0,1)$$

4.Generate random errors according to the normal distribution of normal observations $\varepsilon_i \sim N(0,1)$ and outliers observations:

$$\varepsilon_i \sim N(18,10)$$

The (MCD and MVE) estimators in Tables (1) and (2) achieved more efficiency when estimating the regression parameter β_C for the stratified combined regression estimator to estimate the population mean at all assumed sample sizes (n = 35, 75, 150, 200, 500) and for all proportions. Outliers, because it has a lower MSE value than the traditional covariance matrix estimator, which is sensitive to the presence of outliers, which strongly affect the accuracy of the estimated covariance matrix. We note that the MCD estimator is more efficient than the MVE estimator because it has a lower MSE.

Efficiency values for robust covariance and covariance matrices estimators (MCD, MVE) that were obtained to estimate the classical covariance for all ratios of outliers and assumed sample sizes, were calculated according to formula (13).

3.2 Simulation Results

There are two aspects of our findings:

- **The case of equal sizes of strata:** $(N_1 = N_2 = N_3 = N_4 = N_5 = 140)$
- **The case of unequal sizes of strata:** $(N_1 = 300\ ,\ N_2 = 140\ ,\ N_3 = 95\ ,\ N_4 = 80\ ,\ N_5 = 85)$

In other words, if the layers differ in size, the proportional distribution can be used to maintain the stability of the sampling fraction in every part of the community, if the layer (h) contains (N_h) units, and the sample sizes were calculated according to the proportional distribution formula:

$$n_h = \frac{n\,N_h}{N} \quad , \quad h = 1, 2, \dots, L$$

Table (1)

Values (RE, MSE) to estimate the population mean by combined regression ($\widehat{\bar{Y}}_{lrc}$) when estimating the regression parameter b_c by methods (MCD, MVE, Classic) according to sample sizes (n=35, 75,150,200,500) and the percentages of outliers are (5, 10, 25, 35, and 35) % In the case of equal sizes of strata

n	Methods	The outliers rate is 5%		The outliers rate is 10%		The outliers rate is 25%		The outliers rate is 30%		The outliers rate is 35%	
		MSE	RE	MSE	RE	MSE	RE	MSE	RE	MSE	RE
35	Classic	4.811653	1	5.159734	1	6.063557	1	6.820364	1	6.9881	1
	MCD	4.167706	1.154509	3.697859	1.39533	3.401614	1.782553	4.161787	1.638807	3.778104	1.849631
	MVE	4.230031	1.137498	3.733995	1.381827	3.422285	1.771786	4.164881	1.637589	3.774171	1.851559
75	Classic	1.982465	1	2.032237	1	2.731341	1	3.032213	1	3.068936	1
	MCD	1.642081	1.207288	1.579687	1.286481	1.762943	1.549308	1.639954	1.848962	1.847043	1.66154
	MVE	1.658282	1.195493	1.590765	1.277522	1.762028	1.550112	1.640918	1.847876	1.844492	1.663838
150	Classic	0.89906	1	0.93081	1	1.180187	1	1.348953	1	1.326128	1
	MCD	0.731807	1.228548	0.664413	1.400951	0.744148	1.585957	0.783598	1.721486	0.743696	1.783159
	MVE	0.736658	1.220458	0.699128	1.331387	0.748046	1.577693	0.781334	1.726474	0.743605	1.783377
200	Classic	0.595387	1	0.6895	1	0.875458	1	0.850824	1	0.949779	1
	MCD	0.487447	1.221439	0.551711	1.249749	0.571014	1.533164	0.443856	1.916892	0.574101	1.654376
	MVE	0.495302	1.202069	0.556655	1.238649	0.572068	1.530339	0.444844	1.912635	0.574268	1.653895
500	Classic	0.098733	1	0.089958	1	0.129327	1	0.14272	1	0.153414	1
	MCD	0.072992	1.352655	0.066475	1.353261	0.068537	1.886966	0.083735	1.704425	0.082902	1.850546
	MVE	0.074228	1.330131	0.066991	1.342837	0.068746	1.881229	0.083801	1.703082	0.08301	1.848139

Table (2)

Values (RE, MSE) to estimate the population mean by combined regression ($\widehat{\bar{Y}}_{lrc}$) when estimating the regression parameter b_c by methods (MCD, MVE, Classic) according to sample sizes (n=35, 75,150,200,500) and the percentages of outliers are (5, 10, 25, 35, and 35) %. In the case of unequal sizes of strata

n	Methods	The outliers rate is 5%		The outliers rate is 10%		The outliers rate is 25%		The outliers rate is 30%		The outliers rate is 35%	
		MSE	RE	MSE	RE	MSE	RE	MSE	RE	MSE	RE
35	Classic	3.703451	1	4.462469	1	5.548899	1	5.638562	1	6.229287	1
	MCD	2.637023	1.404406	3.075115	1.451155	2.999428	1.849986	3.054422	1.846032	3.197088	1.948425
	MVE	2.72334	1.359893	3.301191	1.351775	3.204071	1.731828	3.248236	1.735884	3.237289	1.92423
75	Classic	1.774552	1	1.894091	1	2.68928	1	2.543227	1	2.838002	1
	MCD	1.270923	1.39627	1.242852	1.523988	1.69915	1.582721	1.265157	2.010207	1.660658	1.708962
	MVE	1.311529	1.353041	1.218038	1.555034	1.716954	1.566309	1.34278	1.894001	1.692157	1.677151
150	Classic	0.689762	1	0.916689	1	1.100512	1	1.131869	1	1.157773	1
	MCD	0.524054	1.316204	0.677062	1.353922	0.515729	2.133896	0.61471	1.841306	0.594909	1.946135
	MVE	0.545169	1.265226	0.711544	1.28831	0.58355	1.885892	0.676255	1.673731	0.601544	1.924669
200	Classic	0.492493	1	0.566482	1	0.776247	1	0.738385	1	0.810829	1
	MCD	0.381152	1.292117	0.38414	1.474676	0.43356	1.790403	0.420825	1.754613	0.372485	2.17681
	MVE	0.38881	1.266668	0.41439	1.367026	0.456104	1.701908	0.432988	1.705324	0.393287	2.061673
500	Classic	0.07062	1	0.09313	1	0.121836	1	0.129454	1	0.128076	1
	MCD	0.053867	1.311007	0.072194	1.289996	0.065186	1.869052	0.076144	1.700121	0.065243	1.963061
	MVE	0.054235	1.302111	0.068459	1.360376	0.069401	1.755537	0.079647	1.625347	0.067113	1.908364

Table (1) (MSE) was calculated to estimate the community mean by co-regression ($\widehat{\bar{Y}}_{lrc}$)) by estimating the regression parameter b_c. The traditional covariance estimator (Classic), and the MCD estimator are more efficient than the MVE estimator because it has a lower MSE, and for all sample sizes and outliers percentages (10%, 25%, 30%, 35%, 5%), where these values were estimated when applying Formula (13).

From Table (2) (MSE) was calculated to estimate the community mean by covariance ($\widehat{\bar{Y}}_{lrc}$)) by estimating the regression parameter b_c, it was found that the estimators of the robust covariance and covariance matrices (MCD, MVE) are more efficient in terms of having less (MSE) when compared The traditional covariance estimator (Classic), and the MCD estimator is more efficient than the MVE estimator because it has a lower MSE, and for all sample sizes and outliers percentages (10%, 25%, 30%, 35%, 5%), where these values were estimated when applying Formula (13).

4. Conclusions

i. Estimating the population mean by combined regression method using the robust covariance matrix of estimators (MVE, MCD) to estimate the regression parameter is more efficient than using the traditional covariance matrix, when there are outliers in the data sets, for all outliers and assumed sample sizes.

ii. We also conclude that the MCD estimator is more efficient than the MVE estimator because it has the smallest mean square error (MSE) values.

iii. There is an agreement in the results when the sizes of the stratua are equal and when they are of different sizes.

References

1.Alwan, Iqbal Mahmoud, Rasha Adel Saeed, Souad Mahdi Rashid (2011) "Evaluation of stratified random sampling methods in estimating the cultivated area of the coffee crop in Najaf", Department of Statistics, College of Administration and Economics, University of Baghdad, College of Science, Al-Nahrain University.

2.Al-Nasser, Abdul-Majid Hamza, Safaa Younis Al-Safawi (2001). "Theoretical and applied samples", University of Baghdad, University of Mosul, Ministry of Higher Education.

3.AL_Rahman, R. and Mohammad, S., (2022). "Generalized ratio-cum-product type exponential estimation of the population mean in median ranked set sampling". IRAQI JOURNAL OF STATISTICAL SCIENCES, 19(35), pp.83-97.

4.Ahmed, R.A. and Hussein, S.M., (2022). "Generalized modified ratio-cum-product kind exponentially estimator of the populations mean in stratified ranked set sample". International Journal of Nonlinear Analysis and Applications, 13(1), pp.1137-1149.

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

6.Croux, C. and Haesbroeck, G., (1999). "Influence function and efficiency of the minimum covariance determinant scatter matrix estimator". Journal of Multivariate Analysis, 71(2), pp.161-190.

7.Rousseeuw, P.J. and Leroy, A.M., (2005). "Robust regression and outlier detection". John Wiley & sons.

8.Cochran, W.G., (1977)."Sampling techniques". John Wiley & Sons.

9.Rousseeuw, P.J., (1985). "Multivariate estimation with high breakdown point". Mathematical statistics and applications, 8(283-297), p.37.

10. Hawkins, D.M., (1980). "Identification of outliers", (Vol. 11). London: Chapman and Hall.

11. Brus, D.J., (2022)." Spatial Sampling with R". CRC Press.

12. Zaman, T. and Bulut, H., (2019). "Modified ratio estimators using robust regression methods". Communications in Statistics-Theory and Methods, 48(8), pp.2039-2048.

13. Jensen, W.A., Birch, J.B. and Woodall, W.H., (2007). "High breakdown estimation methods for phase I multivariate control charts". Quality and Reliability Engineering International, 23(5), pp.615-629.

14. Hubert, M. and Engelen, S., (2004). "Fast cross validation for high breakdown resampling algorithms".

15. Salibian-Barrera, M. and Yohai, V.J., (2006). "A fast algorithm for S-regression estimates". Journal of computational and Graphical Statistics, 15(2), pp.414-427.

16. Rousseeuw, P.J. and Hubert, M., (2018). "Anomaly detection by robust statistics". Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 8(2), p.e1236.

17. Gupta, S. S., & Singh, H. P. (1983). "Properties of the joint regression estimator in stratified random sampling". Biometrika, 70(1), 269-274.

18. Kim, Y. and Lee, Y. (2005). "Joint regression estimator in stratified random sampling with unequal sample sizes". Journal of Applied Statistics, 32(8), 783-793.

19. Chaudhary, R. and Singh, H. (2014). "On co-regression estimator in stratified random sampling with unequal probability sampling of strata". Statistics in Transition New Series, 15(1), 131-140.

20. Shrestha, S. and Singh, H. (2016). "Combined regression estimator in stratified random sampling with no response". Journal of Applied Statistics, 43(2), 282-293.

تقدير متوسط المجتمع في المعاينة العشوائية الطبقية باستعمال الانحدار المشترك عند وجود القيم الشاذة

مصطفى حبيب مهدي
كلية الادارة والاقتصاد/ جامعة بغداد/ قسم الاحصاء
بغداد، العراق
Mostafa.Habeeb1201a@coadec.uobaghdad.edu.iq

سجى محمد حسين
كلية الادارة والاقتصاد/ جامعة بغداد/ قسم الاحصاء
بغداد، العراق
Saja@coadec.uobaghdad.edu.iq

Received: 28/3/2023 **Accepted: 8/5/2023** **Published: June / 2023**

مستخلص البحث

في هذا البحث ، تم استعمال تقديرات الانحدار المشترك لتقدير متوسط المجتمع في المعاينة العشوائية الطبقية ، وتمت مقارنة تقديرات الانحدار المشترك بتوظيف تقديرات مصفوفة التباين والتباين المشترك الحصينة مع تقديرات الانحدار المشترك بتوظيف تقديرات مصفوفة التباين والتباين المشترك التقليدية عند تقدير معلمة الانحدار ، من خلال معياري الكفاءة (RE) ومتوسط مربعات الخطأ (MSE) .

وجدنا أن التقديرات الحصينة أدت إلى تحسين جودة تقديرات الانحدار المشترك بشكل كبير من خلال تقليل تأثير القيم الشاذة باستعمال تقديرات مصفوفات التباين والتباين المشترك الحصينة(MVE , MCD) عند تقدير معلمة الانحدار .

بالإضافة إلى ذلك، اثبتت نتائج اجراء دراسة المحاكاة ان طريقة اصغر محدد تباين مشترك Minimum covariance determinant (MCD) تكون ذات كفاءة عالية عند جميع احجام العينات (500 , 200 , 150 , 75 , 35=n) ثم تليها طريقة اصغر قطع ناقص Minimum volume ellipsoid (MVE) في التعامل مع القيم الشاذة الموجودة في مجموعة البيانات، حيث تمتلك قيم أقل (MSE) .

نوع البحث: ورقة بحثية .

المصطلحات الرئيسة للبحث: مقدر الانحدار المشترك ($\widehat{\bar{Y}}_{Irc}$) ، القيم الشاذة ، طريقة اصغر محدد تباين مشترك(MCD) ، طريقة اصغر قطع ناقص(MVE) ، الكفاءة ، المعاينة العشوائية الطبقية.