Journal of Economics and Administrative Sciences (JEAS)

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

Comparison of Some Methods for Estimating Parameters of General Linear Model in Presence of Heteroscedastic Problem and High Leverage Points

Qasim Mohammed Saheb
College Of Administration & Economics, University Of Baghdad
qasimalsaheb@gmail.com

Saja Mohammad Hussein
Prof. Dr. of Statistics, College Of Administration & Economics, University Of Baghdad
saja@coadec.uobaghdad.edu.iq

Received: 16/12/2020 **Accepted:31/1/2021** **Published: March/ 2021**

Abstract:

Linear regression is one of the most important statistical tools through which it is possible to know the relationship between the response variable and one variable (or more) of independent variable(s), which is often used in various fields of science. Heteroscedastic is one of linear regression problems, the effect of which leads to inaccurate conclusions. The problem of heteroscedastic may be accompanied by the presence of extreme outliers in the independent variables (High leverage points) (HLPs), the presence of (HLPs) in the data set result unrealistic estimates and misleading inferences. In this paper, we review some of the robust weighted estimation methods that accommodate both Robust and classical methods in the detection of extreme outliers (High leverage points) (HLPs) and determination of weights. The methods include both Diagnostic Robust Generalized Potential Based on Minimum Volume Ellipsoid (DRGP (MVE)), Diagnostic Robust Generalized Potential Based on Minimum Covariance Determinant (DRGP (MCD)), and Diagnostic Robust Generalized Potential Based on Index Set Equality (DRGP (ISE)). The comparison was made according to the standard error criterion of the estimated parameters SE (HC_{4W}) and SE (HC_{5W}) of general linear regression model, for sample sizes (n=60, n=100, n=160), with different degree (severity) of heterogeneity, and contamination percentage (HLPs) are (τ =10%, τ=30%). it was found through comparison that weighted least squares estimation based on the weights of the DRGP (ISE) method are considered the best in estimating the parameters of the multiple linear regression model because they have the lowest standard error values of the estimators (HC_{4W}) and (HC_{5W}) as compared to other methods.

Paper type: Case study

Keywords: Diagnostic Robust Generalized Potential, Robust Heteroscedastic Consistent Covariance Matrix, Masking; Swamping .

1- Introduction:

The homoscedasticity assumption is one of the basic assumptions on which the Ordinary Least Squares (OLS) method depends in estimating the parameters of the linear regression model that are consistent, unbiased, and efficiency. Failure to achieve this assumption result the fact that OLS estimators will not be the best linear unbiased estimate (BLUE) and so the confidence intervals that are set are incorrect. Therefore, do not use the Ordinary least squares (OLS) method when presence of Heteroscedastic problem. outliers observations have an effect even if there is a single observation in the data, its effect may result incorrect conclusions due to distorted estimates of location estimators and dispersion ,and the presence of a single outlier may emptying the properties of the ordinary least squares (OLS).

Detection and processing of outliers observations is essential because their presence in the data set results unrealistic estimates and misleading inferences, the extreme outliers values in the independent variable (HLPs) are considered is one of those outliers observations. classical diagnostic methods fail to correctly detect extreme outliers values (HLPs) due to masking effects, so they are unrealistic methods for determining those values, robust methods are alternative methods as good and effective methods in identifying (HLPs) correctly compared to the methods classical. but robust methods have a tendency to identify more number of outliers in the independent variables (HLPs) and they are not. This reflects the swamping effects, which is also undesirable. DRGP (MVE) method is one of the methods to detect the (HLPs), which is characterized by being an adaptive method that accommodates the two approaches (diagnostic and robust). The robust approach is used in the first stage in identifying suspicious observations as extreme outliers values in the independent variable (HLPs) Then comes the diagnostic methodology as a second stage in confirming from all suspicious observations.

To remedy the problem of Heteroscedastic and presence of extreme outliers values (HLPs) together, we worked on the computation of the Robust Heteroscedastic Consistent Covariance Matrix (RHCCM) for each of the estimators (HC_{4W}) (HC_{5W}) . These estimators included two stages, the first is based on adaptive methods to reduce the effect of (HLPs) through determining the weights and estimation of model parameters, the second stage is the use of a Heteroscedastic Consistent Covariance Matrix (HCCM) in the case of Heteroscedastic to eliminate the effect of the Heteroscedastic problem, the adoption of the standard error criterion in the comparison between the performance of the methods in estimating model parameters.

Among the most important studies that dealt with the problem of heteroscedastic errors and the presence of extreme outlier (HLPs), we mention the study of the researchers (Rousseeuw & Leroy) [8] which included robust estimates of location and dispersion. The study included the detection of leverage points using examples and drawings. .

Furno [4] also adopted the use of robust residuals in finding the Heteroscedastic Consistent Covariance Matrix (HCCM) when there are outliers in the independent variables (HLPs), which required the computation of the robust weights to reduce the effect of the outliers. The idea of using robust

residuals is to reduce the bias of HCCM estimators and to obtain consistent and robust of Covariance Matrix (RHCCM) estimators.

In 2018, the researchers (Midi) and others [7] presented a comparison study between some of the robust weighted least squares methods in estimating model parameters when there is a problem of heteroscedastic errors and the presence of extreme outliers in the independent variables (high leverage points) (HLPs) by finding the Robust Heteroscedastic Consistent Covariance Matrix (RHCCM) that included both of the estimators (HC_{4W}) (HC_{5W}) , based on Furno's method of adding the robust weight matrix to the estimators, which were found based on the RMD (MVE) and DRGP (ISE) as detection methods for HLPs . The results showed that the weighted least squares method which is based on the DRGP (ISE) method is the best compared to other methods and in various percentage of outliers in the independent variables (HLPs).

This paper a review of some robust weighted methods on the basis of measures of detecting extreme outliers values in the independent variable (high leverge points) (HLPs) and their use in the weighted least squares method to estimate the parameters of the general linear regression model when there is a problem of heteroscedastic and outliers in the independent variable (high leverge points). The robust estimates of the parameters of the general linear regression model were compared through the (SE) criterion of estimators (HC_{4W}) (HC_{5W}).

2- Theoretical Section

Linear regression is a functional relationship written in the form of a linear equation, and Considered one of its uses is to explain a variable that called the response variable through one or more of the independent variables , it is calculated according to the following formula:

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip} + \varepsilon_i \quad , i = 1, 2 \dots, n \qquad (1)$$

Where

y_i : observations of the Response variable

$x_{i1}, x_{i2}, \dots, x_{ip}$: observations of the independent variables

$\beta_0, \beta_1, \beta_2, \dots, \beta_p$: parameters of the linear regression model

ε_i : The random error is a random variable that is assumed to be normally distributed with a mean zero and constant variance σ^2

The assumption that the random error limit is constant, or what is known as the homogeneity hypothesis is achieved if the observations were drawn from identical populations and had the same variance, i.e.

$$\sigma^2{}_U = \sigma^2{}_1 = \sigma^2{}_2 = \cdots \dots . = \sigma^2{}_n \qquad (2)$$

the general formula

$$E(\varepsilon\varepsilon') = \begin{bmatrix} \sigma^2{}_1 & 0 & \cdots & 0 \\ 0 & \sigma^2{}_2 & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \sigma^2{}_n \end{bmatrix} = \sigma^2 \begin{bmatrix} 1 & 0 & \cdots & 0 \\ 0 & 1 & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 1 \end{bmatrix} = \sigma^2 I_n$$

But when this hypothesis is not achieve, the problem of heteroscedastic errors appears, or what is known the error limit is inconstant, we often find it in cross section data .Weighted least squares (WLS) method is used in the process of estimating the parameters of the linear regression model as a method of dealing

with the problem of heteroscedastic by specifying weights (wi) that make the variances of errors equal, which is calculated by the following formula:

$$\underline{\widehat{B}}_{WLS} = (X'W^{-1}X)^{-1}X'W^{-1}Y \qquad (3)$$

This method is known as the Generalized Least Squares Method, because of its dependence on weights (wi) it is known as the Weighted Least Squares Method (WLS).

The covariance matrix for the estimated parameters is computed in the following :

$$var - cov(\widehat{B}_{WLS}) = \sigma^2(X'W^{-1}X)^{-1} \qquad (4)$$

3- Estimation the parameters of the linear regression model

When there is a problem of heteroscedastic the weighted least squares method is used to eliminate the effect of this problem, which requires finding a matrix of weights in estimating the parameters of the linear regression model. The presence of extreme outliers in the independent variable (HLPs) in addition to the presence of the problem of heteroscedastic. We will obtain misleading results due to the effect of these values when using the classical methods. These weights were found by some adaptive methods that accommodate the two approaches (diagnostic and robust), in order to reduce the swamping effects of the robust methods when the extreme outliers are present in the independent variables (HLPs) and reduce their effect when estimating the parameters of the General linear regression model , the methods are :

3-1 Diagnostic Robust Generalized Potential Based on Minimum Volume Ellipsoid (DRGP(MVE))

Classical diagnostic methods are affected by the effects of masking when used in detecting outliers in the independent variables (HLPs), which makes them an unrealistic method of identifying those values. The robust methods are alternative methods as good and effective methods for identifying (HLPs) correctly compared to the classical methods, but the robust methods have the tendency to specify a greater number of outliers in the independent variables (HLPs) which are not, this reflects the effects of swamping and this is also undesirable. [5]

(Habshah) and others [6] proposed Diagnostic Robust Generalized Potential Based on Minimum Volume Ellipsoid (DRGP(MVE)) which is an adaptive method that accommodates the two approaches ,The robust approach is used in the first stage in identifying suspicious observations as extreme outliers in the independent variable (HLPs) Then the diagnostic methodology comes as a second stage in confirming from all suspicious observations.

Actually finding the (MVE) estimators can be very difficult in practice by the use of combination when sample size (n) and the variables number (p) increase , because of increasing the required computational effort dramatically which it takes a long time. Rousseeuw and Leroy [8] proposed an approximate method to find the (MVE) estimators involving the use of a subsampling algorithm through determine number of subsamples from among all the subsamples That's drawn.[8 pp. 260]

The Minimum Volume Ellipsoid estimators are used in the first stage by applying robust Mahalanobis distance RMD (MVE) as a robust method for detecting outliers in the independent variables (HLPs). it can be summarized in a number of steps:

1.draw all possible subsamples J of size (P + 1) from observations according to combinations:

$$\binom{n}{P+1} \qquad J = 1, 2, \dots\dots, P+1 \qquad (5)$$

2.Calculation of Mahalanobis distances for the observations of the estimators of location and dispersion for each subsamples according to the following formula:

$$MD_i = \sqrt{\left(x_i - \underline{\hat{\mu}_i}\right)\ (cv)^{-1}\left(x_i - \underline{\hat{\mu}_i}\right)'} \qquad (6)$$

Where

$\underline{\hat{\mu}_i}$ **The position estimator, is the vector of the arithmetic mean for each subsample**

cv **The dispersion estimator, is the variance and covariance matrix for each subsample**

3.Calculate the Minimum Volume Ellipsoid for all subsamples according to the following formula :

$$\text{volume} = \sqrt{det\ (m_j^2\ cv)} = m_j^p\ det(cv) \qquad (7)$$

where

m_j **The maximization factor that is used to maximize the subsample size to contain approximately half of the observations**

4.The subsample that has the Minimum Volume Ellipsoid from all the subsamples according to Equation (7) is the Optimum subsample.

5.Calculate the robust Mahalanobis distances RMD (MCD) for all sample observations by replace the mean estimator of the Minimum Volume Ellipsoid and the covariance matrix estimator after multiplying it by a suitable factor instead of the arithmetic mean and the covariance matrix in the classical Mahalanobis distances formula.

6.Determine the outliers by testing them with an appropriate cut-off point, on this basis any robust Mahalanobis distances greater than the cut-off value is considered an extreme outlier in the independent variable (HLP). The cut-off point for robust Mahalanobis distances is defined as (cd) and is calculated as follows : [7][9][10]

$$cd = median(RMD_i) + 3MAD(RMD_i) \qquad (8)$$

Where

$median(RMD_i)$ **: Median value of robust Mahalanobis distances**

$MAD(RMD_i)$ **: The median of absolute deviations Mahalanobis distances from the median**

$$MAD(RMD_i) = median\left|(RMD_i) - median(RMD_i)\right| \qquad (9)$$

Any observation whose value exceeds the cut-off value is considered to be a suspicious value (HLP) and is placed in group D or what is also known as a deleted set. The remaining group of observations represents group R (remaining group), which It has observations size (n-d) .After identifying both groups (D) and (R), the first stage ends and the diagnostic methodology begins to check all observations of group D. Without losing generality, we assume that we have the observations of the D group, which are the last rows of each of the variables X and Y, to calculate the hat matrix we use the following formula:

$$\mathbf{H} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}' \qquad (10)$$

Depending on both the group of deleted observations (D) and the remaining group (R), it is possible to find the elements of $h_i^{(-D)}$ we use the following formula:

$$h_i^{(-D)} = x_i'(X_R'X_R)^{-1}x_i \qquad i = 1, 2, \dots\dots\dots, n \qquad (11)$$

Where $h_i^{(-D)}$ represents the diagonal elements of the observations in the matrix $X(X_R'X_R)^{-1}X'$, it represents the diagonal elements of the hat matrix without a group (D). Depending on the generalized potential equation, we find the potential value of all observations in both group (R) and group (D), which is known as the following: [5]

$$p_i^* = \begin{cases} \dfrac{h_i^{(-D)}}{1 - h_i^{(-D)}} & for\ i \in R \\ h_i^{(-D)}, & for\ i \in D \end{cases} \qquad (12)$$

Then the outliers in the independent variables (HLPs) are determined by comparing the potential value p_i^* with the cut-off point :

$$Cd_M = \mathrm{median}\,(p_i^*) + CMAD\,(p_i^*) \qquad (13)$$

if all the values in group D are greater than the cut off point in equation (13) then all those values will be declared outliers in the independent variables (HLPs), but if all values are not greater than the cut-off point, those values will be returned to the group R in sequence (the value with the lowest potential value p_i^* is returned at the beginning), then the potential values p_i^* are calculated, this process continues until all the values in the group (D) are (HLPs). [6]

The diagonal elements of the robust weights matrix can be found according to the following formula:

$$w_{iDM} = min\left(1, \frac{cd_M}{\widehat{p_i}}\right) \qquad (14)$$

The values which detected as outliers in the independent variables (HLPs) in the final group (D) will take the weight ($\frac{cd_M}{\widehat{p_i}}$), while the rest of the usual observations take weight (1). [7]

3-2 Diagnostic Robust Generalized Potential Based on Minimum Covariance Determinant (DRGP(MCD))

Diagnostic Robust Generalized Potential Based on Minimum Covariance Determinant (DRGP(MCD)) is the second method that will be employed in finding robust weights used in the formula for estimating the parameters of the linear regression model according to the weighted least squares method when presence problem of heteroscedastic and outliers in the independent variable (HLPs) together. This method also works in two stages, as is the case in the

(DRGP(MVE)) method . The first stage begins in detecting the values of (HLPs) through the use Robust Mahalanobis Distance. Finding estimators (MCD) and display the work of RMD (MCD) in detecting outliers observations in the variables X can be summarized in the following Algorithm:

1.draw all possible subsamples J of size (h) from observations according to combinations:

$$\binom{n}{h} = \frac{n!}{h!\,(n-h)!} \qquad J = 1, 2, \dots\dots, h \qquad (15)$$

Where h indicate to half the size of the data, which equal

$$h = \frac{n+p+1}{2} \qquad (16)$$

2.Finding the best subsample J, which is the subsample that has the Minimum Covariance Determinant among all subsamples :

$$\mathcal{J} = arg_J\ min\left|cv\right| \qquad (17)$$

3.considered the estimators of the best subsample for each of the arithmetic mean vector and the covariance matrix after multiplying them by suitable factor are estimators (MCD) for the location and dispersion, then we work to find the square Mahalanobis distances for all the sample observations based on those estimators according to the following formula :

$$RMD_{iMCD}{}^2 = \left(x_i - \underline{\hat{\mu}}_{j\,MCD}\right)(cv_{MCD})^{-1}\left(x_i - \underline{\hat{\mu}}_{j\,MCD}\right)' \qquad (18)$$

4.To find outliers extreme (HLPS) we will test the computed robust Mahalanobis distances according to the cut-off point used in the DRGP (MVE) method in equation (8). Any robust distance greater than the cut-off value will be considered an extreme outliers value (HLP), otherwise it is a normality observation. [10]

The values of the Robust Mahalanobis Distance that are exceed the cut-off point value are known as the suspicious values as extreme outliers in the independent variable (HLPs) and are placed in a group D, while the rest of the normality values are placed in the R group, this ends the first stage. The second stage in the (DRGP (MCD)) method is the same as the second stage in the (DRGP (MVE)) method, even in the cut-off point that plays the most prominent role in determining the robust weights.

3-3 Diagnostic Robust Generalized Potential Based on Index Set Equality (DRGP(ISE))

Both the DRGP (MVE) method and the DRGP (MCD) method in the first stage depend on the Robust Mahalanobis Distance based on each of the estimators (MVE) (MCD) in detecting (HLPs), which requires a large computational effort and takes a long time. (Lim and Midi) [5] They proposed to use the Diagnostic Generalized potential method (DRGP) based on the Index Set Equality (ISE) when finding the robust estimator of both location and dispersion. Let's display the observations of the independent variables in the form of a row vector as in the following formula:

$$x_i = (1, x_{i1}, x_{i2}, \dots\dots\dots, x_{ik}) = (1, R_i) \qquad (19)$$

Where

$$R_i = (x_{i1}, x_{i2}, \dots\dots\dots, x_{ik}) \qquad (20)$$

Then let us refer to the old indexing group (ISold), which are arranged in ascending order as follows:

$$IS_{old} = \{\pi^{old}{}_1, \pi^{old}{}_2, \dots\dots, \pi^{old}{}_h\} \quad (21)$$

This index set corresponds to the elements of the half subset when the squared Mahalanobis distance are arranged in ascending of the half subset.[5][7]

The first stage can be summarized in identifying the suspicious observations as extreme outliers in the independent variable (HLPs) using RMD (ISE) according to the following algorithm:

1.Selecting a subset denoted by the symbol H_{old} and which contains h observations.

2.Calculate the arithmetic mean vector and the covariance matrix of the subset H_{old}

3.Calculate the square Mahalanobis distance for all observations in H_{old} according to the following formula:

$$D^2{}_{old} = \left(x_i - \underline{\hat{\mu}_i}_{Hold}\right) (cv_{Hold})^{-1} \left(x_i - \underline{\hat{\mu}_i}_{Hold}\right)' \quad (22)$$

4.Arrange the square Mahalanobis distance in equation (22) in ascending order as shown below.

$$d^2{}_{old}(\pi(1)) \leq d^2{}_{old}(\pi(2)) \leq \cdots \leq d^2{}_{old}(\pi(h)) \quad (23)$$

5.Construct a new subset known as H_{new} according to the following formula:

$$H_{new} = \{R_{\pi(1)}, R_{\pi(2)}, \dots\dots, R_{\pi(h)}\} \quad (24)$$

This set corresponds to the elements of the new index set

$$IS_{new} = \{\pi^{new}{}_1, \pi^{new}{}_2, \dots\dots, \pi^{new}{}_h\} \quad (25)$$

6.If $(IS_{old} \neq IS_{new})$ let $(H_{old} := H_{new})$ and work on the calculation of $CV_{H(new)}$ and let $(CV_{H(old)} := CV_{H(new)})$ Then we calculate $\underline{\hat{\mu}_i}_{Hnew}$ We also let that $(\underline{\hat{\mu}_i}_{H(old)} := \underline{\hat{\mu}_i}_{H(new)})$ Then repeat steps (3-7) to get $(IS_{old} = IS_{new})$ when this is achieved both of them $\underline{\hat{\mu}_i}_{H(new)}$ and $CV_{H(new)}$ are the estimators position and scattering of the (ISE) method, substituting these estimators into the Mahalanobis distances formula to find the robust Mahalanobis distances for all sample observations.

To detect the extreme outliers in the independent variable (HLPs) in the first stage will be depend on the cut-off point in equation (8). The robust Mahalanobis distances that exceed the cut-off value are placed in a group (D), while the rest of the normal values are placed in the group R, [7]

The second stage of the robust Diagnostic Generalized potential Method (DRGP (ISE)) works to check from the extreme outliers (HLPs) in group D that were detected in the first stage, which is the same steps of the second stage in the DRGP (MVE) method. Except in the cut-off point, in this method we used another cut-off point which is known as the following formula:

$$Cd_I = median(\hat{p}_i) + 3\, Q_n(\hat{p}_i) \quad (26)$$

Where

Q_n it is a robust estimator that has a breakdown point of up to (50)% and represents the pairwise order statistic for all distance, which increases the accuracy of the cut-off point in determining (HLPs) is known as the following formula:

$$Q_n = \{ C \mid x_i - x_j \mid i < j \}_{(k)} \quad (27)$$

Where *c* = 2.2219 this value will provides Q*n* a consistent estimator for Gaussian data

$$k = \binom{h}{2} \approx \binom{h}{2}/4$$

Depending on each group(R, D) the process is repeated until it is realized that all values in group D are extreme outliers (HLPs). The diagonal elements of the robust weights matrix are calculated according to the following formula: [5]

$$w_{iDI} = min\left(1, \frac{cd_I}{\widehat{p_i}}\right) \quad (28)$$

observations corresponding to HLPs will take weight ($\frac{cd_I}{\widehat{p_i}}$), the rest of the normality values will take weight (1) .

4- Robust Heteroscedastic Consistent Covariance Matrix Estimate for Estimate of the Regression linear model parameters (RHCCM)

The HC_0 estimator proposed by white [11] is the first consistent estimator for the covariance matrix under both cases of homoscedastic and heteroscedastic of unknown form, which is based on the OLS estimates when there is a problem of heteroscedastic. estimators HC_1, HC_2, HC_3, are proposed and then proposed another new estimator known as HC_4 by (Cribari Neto) [3] which depend on the estimator HC_3, except that The HC_4 takes into account the ratio between the measurement of the (hi) and its mean . The HC_5 estimator is another proposal by Cribari-Neto and others [2], this estimator takes into account the maximum leverage (hi) in addition to the features of the estimator HC_4. Presence of the problem of heteroscedastic and extreme outliers (HLPs) together makes estimates of OLS parameters biased, thus the inference becomes unrealistic. (Furno) [4] suggested the Robust Heteroscedastic Consistent Covariance Matrix (RHCCM) in the case of heteroscedastic and presence the extreme outliers in the independent variable (HLPs), Weighted least square residuals (WLS) is adopted instead of residuals OLS in estimators HCCM. [7]

this paper depend on both the estimators HC_5 and HC_4, because they take into account the presence of extreme outliers in the independent variable (HLPs) and based on Furno's method in computing the Robust Heteroscedastic Consistent Covariance Matrix For the parameters of the estimated model according to the following formulas and Respectively:

$$HC_{4W} = (X'WX)^{-1}X'W\widehat{\psi}_{4w}\,WX\,(X'WX)^{-1} \quad (29)$$

where W is a diagonal square matrix with diagonal elements wi

$$\widehat{\psi}_4 = diag\left\{\frac{\hat{e}_i^{\,2}}{(1-h_i^{\,*})^{\delta i}}\right\} \qquad i = 1, 2, \dots\dots n$$

$$\delta i = min\left\{4, \frac{h_i^{\,*}}{\bar{h}^*}\right\} \qquad , \bar{h}^* = \frac{\sum h_i^{\,*}}{n}$$

$$HC_{5W} = (X'WX)^{-1}X'W\widehat{\psi}_{5w}\,WX\,(X'WX)^{-1} \quad (30)$$

$$\widehat{\psi}_5 = diag\left\{\frac{\hat{e}_i^{\,2}}{\sqrt{(1-h_i^{\,*})^{\alpha i}}}\right\} \qquad i = 1, 2, \dots\dots n$$

$$\alpha i = min\left\{\frac{h_i^*}{\bar{h}^*}, max\left\{4, \frac{kh^*_{max}}{\bar{h}^*}\right\}\right\}$$

$h_i^* = \sqrt{W}\, x_i\, (X'WX)^{-1} x_i' \sqrt{W}$

h_i^* : the diagonal elements of the weighted hat matrix

$$H_W = \sqrt{W}X\, (X'WX)^{-1} X'\sqrt{W} \qquad (31)$$

It is adopted as being equal to (0.7). [2] where k is a constant $(0 < k < 1)$

5- Experimental section

The experimental section included generating data that was written in the programming language (R) using the simulation method (Monte Carlo) with iterations (10,000) for each experiment, that each experiment is a random process independent of any other experiment, for sample sizes (n = 60, n = 100, n = 160) and with different values of the degree Heteroscedastic and with percentage of outliers (HLPs) equal to (τ = 10%, τ = 30%), compared were among methods DRGP (MVE) , DRGP (MCD) and (DRGP (ISE)) in estimating the model parameters according to the criterion of standard errors (SE) through the Robust Heteroscedastic Consistent Covariance Matrix (RHCCM) and for each of the estimators (HC_{4W}) (HC_{5W}).

The simulation involved a number of steps in generating the data, which are :

1.Generate the independent variables (x (1), x (2), x (3), x (4)) according to the standard normal distribution and according to the following formula :

$X \sim N(0, 1)$

2.Assuming the parameters of the original model are equal to the following value :

$B_0 = 1.5$, $B_1 = B_4 = 1$, $B_2 = 0.6$, $B_3 = 0.4$

3.Generating random errors according to the normal distribution according to the following formula :

$\epsilon_i \sim (o, \sigma^2{}_i)$, $i = 1, 2, \dots, n$

Where

$\sigma^2{}_i = \exp\{c * x_{i1}\}$

Where ($\sigma^2{}_i$) represents generating the Heteroscedastic function. The degree (severity) of the Heteroscedastic depends on the value of the perturbation constant (C). If the value of (C=0), this leads to homoscedastic, with an increase in the value of (C) the severity of the Heteroscedastic is increases.in this paper will be two severity of the Heteroscedastic is low and high depending on the value of (C). When we adopt the value of (C = 0.20) we obtain a (low λ), but if the value of (C = 0.40) we obtain a (high λ). The severity of the Heteroscedastic is calculated according to the following:

$$\lambda = \frac{\max(\sigma^2{}_i)}{min(\sigma^2{}_i)} \qquad , i = 1, 2 \dots, n \qquad (32)$$

4.calculate the response variable by multiplying the matrix of independent variables with vector the parameters, adding the value of the random error

5.In order to obtain (HLPs) in the data, some observations in the independent variables were randomly replaced with extreme outliers (HLPs) values with specific percentage (30%, 10%) of the total sample size, these values follow the normal distribution according to the formula The following:

$X \sim N(15, 1)$

5-1 Results of the Simulation Experiments

The results of the simulation experiments included a number of tables:

Table (1) shows the estimates of the parameters of the general linear regression model for the robust weighted methods and standard error (SE) for the estimators (HC_{4W}) (HC_{5W}) when (n = 60) and the percentage of extreme outliers values in the independent variable (HLPs = 10%, HLPs = 30%)

HLP	Method	(low λ)			(high λ)		
		Estimate	Se.HC_{W4}	Se.HC_{W5}	Estimate	Se.HC_{W4}	Se.HC_{W5}
10%	b0.DRGP(MVE)	1.499268	0.136275*	0.134079	1.500738	0.141205*	0.138902
	b1.DRGP(MVE)	1.000135	0.150945*	0.14495	0.998447	0.166513*	0.161455
	b2.DRGP(MVE)	0.599795	0.064458	0.050103	0.600595	0.066205*	0.054799
	b3.DRGP(MVE)	0.39819	0.153991*	0.147817	0.396743	0.140628*	0.135057
	b4.DRGP(MVE)	0.999647	0.057976	0.046977	1.000208	0.053934	0.045093
	b0.DRGP(MCD)	1.49925	0.136312	0.134072	1.500829	0.141409	0.139086
	b1.DRGP(MCD)	1.000087	0.151505	0.145199	0.998504	0.166522	0.161361
	b2.DRGP(MCD)	0.599796	0.064449*	0.050107	0.600582	0.066359	0.054857
	b3.DRGP(MCD)	0.398218	0.154271	0.148035	0.396641	0.141148	0.135493
	b4.DRGP(MCD)	0.99965	0.057968*	0.04698	1.000204	0.053935	0.045094
	b0.DRGP(ISE)	1.499246	0.136315	0.134008*	1.500757	0.141387	0.138857*
	b1.DRGP(ISE)	1.000101	0.151447	0.144711*	0.998603	0.167212	0.161097*
	b2.DRGP(ISE)	0.599791	0.065019	0.049475*	0.600533	0.06766	0.054422*
	b3.DRGP(ISE)	0.398255	0.15442	0.147163*	0.396788	0.141277	0.134578*
	b4.DRGP(ISE)	0.999663	0.058495	0.046513*	1.000185	0.053601*	0.043614*
30%	b0.DRGP(MVE)	1.501132	0.156819	0.154127*	1.49968	0.160042	0.158145
	b1.DRGP(MVE)	1.00166	0.132927	0.126552*	1.000692	0.142489*	0.137946*
	b2.DRGP(MVE)	0.599703	0.034526	0.030267*	0.600211	0.035937*	0.032084
	b3.DRGP(MVE)	0.400021	0.034509	0.030481*	0.399904	0.032791	0.030375
	b4.DRGP(MVE)	0.999979	0.030219	0.027424*	0.999775	0.032482*	0.029096
	b0.DRGP(MCD)	1.501132	0.156819	0.154127*	1.49968	0.160042	0.158145
	b1.DRGP(MCD)	1.00166	0.132927	0.126552*	1.000692	0.142489*	0.137946*
	b2.DRGP(MCD)	0.599703	0.034526	0.030267*	0.600211	0.035937*	0.032084
	b3.DRGP(MCD)	0.400021	0.034509	0.030481*	0.399904	0.032791	0.030375
	b4.DRGP(MCD)	0.999979	0.030219	0.027424*	0.999775	0.032482*	0.029096
	b0.DRGP(ISE)	1.501129	0.156175*	0.154435	1.499755	0.15993*	0.157954*
	b1.DRGP(ISE)	1.001767	0.130608*	0.126582	1.00086	0.143515	0.13799
	b2.DRGP(ISE)	0.599718	0.034119*	0.0303	0.600196	0.035995	0.03184*
	b3.DRGP(ISE)	0.400004	0.034235*	0.030987	0.399885	0.03279*	0.030294*
	b4.DRGP(ISE)	1.000011	0.030077*	0.027583	0.999788	0.032505	0.028846*

The sign (*) denotes the lowest value of the standard error (SE) of the estimated parameter compared to its counterparts from other methods.

Table (2) shows the estimates of the parameters of the general linear regression model for the robust weighted methods and standard error (SE) for the estimators (HC_{4W}) (HC_{5W}) when (n = 100) and the percentage of extreme outliers values in the independent variable (HLPs = 10%, HLPs = 30%)

HLP	Method	(low λ)			(high λ)		
		Estimate	Se.HC_{W4}	Se.HC_{W5}	Estimate	Se.HC_{W4}	Se.HC_{W5}
10%	b0.DRGP(MVE)	1.500966	0.114725*	0.113489	1.254365	0.098899	0.097484
	b1.DRGP(MVE)	1.000899	0.112125	0.110333	0.836151	0.098534	0.096065
	b2.DRGP(MVE)	0.600014	0.034522	0.030852	0.501163	0.032301	0.028012*
	b3.DRGP(MVE)	0.398611	0.11267	0.110746	0.333996	0.085537*	0.08343*
	b4.DRGP(MVE)	1.000275	0.038671	0.034307	0.83542	0.033598	0.029163*
	b0.DRGP(MCD)	1.500965	0.114734	0.11348*	1.254365	0.098899	0.097484
	b1.DRGP(MCD)	1.000913	0.112029*	0.110297*	0.836151	0.098534	0.096065
	b2.DRGP(MCD)	0.600013	0.034524	0.030854	0.501163	0.032301	0.028012*
	b3.DRGP(MCD)	0.398628	0.112608*	0.110666	0.333996	0.085537*	0.08343*
	b4.DRGP(MCD)	1.000276	0.038665	0.034294	0.83542	0.033598	0.029163*
	b0.DRGP(ISE)	1.50097	0.114779	0.113495	1.254366	0.098854*	0.09745*
	b1.DRGP(ISE)	1.000891	0.112383	0.110476	0.83616	0.098386*	0.095989*
	b2.DRGP(ISE)	0.600036	0.034303*	0.030191*	0.501166	0.032237*	0.028141
	b3.DRGP(ISE)	0.398614	0.112713	0.110635*	0.334017	0.085698	0.083698
	b4.DRGP(ISE)	1.000262	0.038627*	0.033772*	0.835422	0.033521*	0.029248
30%	b0.DRGP(MVE)	1.500478	0.115961	0.115323	1.500928	0.120273*	0.119456*
	b1.DRGP(MVE)	1.000873	0.102757	0.101268	1.001498	0.116977	0.115102
	b2.DRGP(MVE)	0.5998	0.024204	0.022907	0.600365	0.023411	0.02205
	b3.DRGP(MVE)	0.399993	0.025419	0.023957	0.3998	0.022833	0.021724
	b4.DRGP(MVE)	0.999897	0.02194	0.021059	1.000047	0.023106	0.02194
	b0.DRGP(MCD)	1.500478	0.115961	0.115323	1.500928	0.120273*	0.119456*
	b1.DRGP(MCD)	1.000873	0.102757	0.101268	1.001498	0.116977	0.115102
	b2.DRGP(MCD)	0.5998	0.024204	0.022907	0.600365	0.023411	0.02205
	b3.DRGP(MCD)	0.399993	0.025419	0.023957	0.3998	0.022833	0.021724
	b4.DRGP(MCD)	0.999897	0.02194	0.021059	1.000047	0.023106	0.02194
	b0.DRGP(ISE)	1.500473	0.11579*	0.115129*	1.50093	0.120494	0.119607
	b1.DRGP(ISE)	1.000921	0.10184*	0.100125*	1.001409	0.116544*	0.114416*
	b2.DRGP(ISE)	0.599797	0.024021*	0.022629*	0.600368	0.023343*	0.021927*
	b3.DRGP(ISE)	0.400016	0.025266*	0.023702*	0.399788	0.022821*	0.021676*
	b4.DRGP(ISE)	0.999903	0.021824*	0.020895*	1.000051	0.023044*	0.021816*

The sign (*) denotes the lowest value of the standard error (SE) of the estimated parameter compared to its counterparts from other methods.

Table (3) shows the estimates of the parameters of the general linear regression model for the robust weighted methods and standard error (SE) for the estimators (HC_{4W}) (HC_{5W}) when (n = 160) and the percentage of extreme outliers values in the independent variable (HLPs = 10%, HLPs = 30%)

HLP	Method	(low λ)			(high λ)		
		Estimate	Se.HC_{W4}	Se.HC_{W5}	Estimate	Se.HC_{W4}	Se.HC_{W5}
10%	b0.DRGP(MVE)	1.501635	0.085173	0.084624*	1.500037	0.085303	0.084856
	b1.DRGP(MVE)	1.001859	0.080521	0.079286	0.999535	0.096145	0.094864
	b2.DRGP(MVE)	0.599636	0.027233*	0.024217*	0.599889	0.026797	0.025077
	b3.DRGP(MVE)	0.399786	0.080766	0.078012*	0.400109	0.087108	0.08621
	b4.DRGP(MVE)	0.999547	0.028542*	0.0253*	0.999683	0.029841	0.027673
	b0.DRGP(MCD)	1.501635	0.085173	0.084624*	1.500038	0.085299*	0.084854
	b1.DRGP(MCD)	1.001859	0.080521	0.079286	0.99954	0.095947	0.094693
	b2.DRGP(MCD)	0.599636	0.027233*	0.024217*	0.59989	0.026778	0.025057
	b3.DRGP(MCD)	0.399786	0.080766	0.078012*	0.400111	0.087089	0.086194
	b4.DRGP(MCD)	0.999547	0.028542*	0.0253*	0.999683	0.029836	0.027667
	b0.DRGP(ISE)	1.501587	0.085149*	0.084656	1.500028	0.0853	0.084848*
	b1.DRGP(ISE)	1.001831	0.079938*	0.079083*	0.99953	0.09583*	0.094543*
	b2.DRGP(ISE)	0.59962	0.027365	0.025174	0.599887	0.026651*	0.02486*
	b3.DRGP(ISE)	0.399631	0.080072*	0.079137	0.400092	0.086889*	0.085954*
	b4.DRGP(ISE)	0.999267	0.031422	0.028453	0.999694	0.029754*	0.02751*
30%	b0.DRGP(MVE)	1.498911	0.090919*	0.090534*	1.500836	0.093829	0.093457
	b1.DRGP(MVE)	0.99934	0.08161*	0.079977*	1.00096	0.093732	0.09271
	b2.DRGP(MVE)	0.600191	0.017759*	0.017034*	0.600198	0.018515	0.017843
	b3.DRGP(MVE)	0.400161	0.018546*	0.017654*	0.400094	0.018972	0.018441
	b4.DRGP(MVE)	1.000017	0.018793*	0.017939*	0.999922	0.018182	0.017594
	b0.DRGP(MCD)	1.498911	0.090919*	0.090534*	1.500836	0.093829	0.093457
	b1.DRGP(MCD)	0.99934	0.08161*	0.079977*	1.00096	0.093732	0.09271
	b2.DRGP(MCD)	0.600191	0.017759*	0.017034*	0.600198	0.018515	0.017843
	b3.DRGP(MCD)	0.400161	0.018546*	0.017654*	0.400094	0.018972	0.018441
	b4.DRGP(MCD)	1.000017	0.018793*	0.017939*	0.999922	0.018182	0.017594
	b0.DRGP(ISE)	1.498935	0.091046	0.090734	1.500851	0.09378*	0.093388*
	b1.DRGP(ISE)	0.999266	0.083912	0.082953	1.001062	0.093386*	0.09219*
	b2.DRGP(ISE)	0.600177	0.017993	0.017369	0.600198	0.018413*	0.017702*
	b3.DRGP(ISE)	0.400157	0.019018	0.018261	0.400098	0.018894*	0.018332*
	b4.DRGP(ISE)	1.000079	0.019134	0.018408	0.999924	0.018058*	0.017439*

The sign (*) denotes the lowest value of the standard error (SE) of the estimated parameter compared to its counterparts from other methods.

5-2 Discuss the results of the simulation experiments

Through the results of simulation experiments in Tables (1) (2) (3), we note the following:

1.When the percentage of (HLPs = 10%) and with the different severity of the Heteroscedastic and in all sample sizes (n = 60, n = 100, n = 160), the DRGP (ISE) method is considered the best in estimating the model parameters according to the standard error (SE) of the estimators (HC_{4W}) and (HC_{5W}).

2.When the percentage of (HLPs = 30%), we find that the (DRGP (ISE)) method is the best in estimating the model parameters according to the standard error criterion (SE) of the estimators (HC_{4W}) and (HC_{5W}) in the sample sizes (n = 60, n = 100), but when it is (n = 160) we find that all methods have the same preference.
3.Through all the simulation results, it was found that the DRGP (ISE) method is the best in detecting extreme outliers (HLPs) and reducing its effect on estimating the parameters of the general linear regression model with different sample sizes, percentage (HLPs), and severity of the Heteroscedastic .

6- Conclusions:

1. The DRGP (ISE) method is the more efficient in estimating the parameters of the linear regression model because it achieved the lowest standard error (SE) of estimators (HC_{4W}) and (HC_{5W}) in sample sizes (n = 60, n = 100) , in different percentage of extreme outliers values (HLPs) in the data , and severity difference of Heteroscedastic.
2. We conclude from simulation experiments that all DRGP methods have the same preference to reduce the influence of extreme outliers values (HLPs) in estimating the parameters of the linear regression model when increase the sample size (n = 160) and the contamination percentage (HLPs=30%).

7- Further Work

1.Using the DRGP (ISE) method in estimating the parameters of the regression model when there are extreme outliers in the independent variable (HLPs).
2.We recommend the adoption of (DRGP (ISE)) method in determining the weights, for both the estimator (HC_{4W}) and (HC_{5W}), for different sample sizes, with difference the percentage of the extreme outliers (HLPs) in the data and severity of the Heteroscedastic .

References

1.Kazim ,A,H,. ALDoulami ,M,M,.(1988) "Introduction to Linear Regression Analysis" University of Mosul - Directorate of Books for Printing and Publishing.
2.Cribari-Neto, F., Souza, T.C.and Vasconcellos, K.L.P. "Inference under heteroskedasticity and leveraged data" . Theory Methods , vol (36) ,pp. 1877–1888 (2007). Errata :vol (37),pp. 3329–3330 (2008)
3.Cribari-Neto, F. (2004). "Asymptotic inference under heteroskedasticity of unknown form" .Computational Statistics and Data Analysis, vol (45), pp. 215–233.
4.Furno, M. (1996) " Small sample behavior of a robust heteroskedasticity consistent covariance matrix estimator" Journal of Statistical Computation and Simulation, vol (54), pp. 115-128.
5.Lim, H.A., Midi , H. (2016) "Diagnostic Robust Generalized Potential Based on Index Set Equality (DRGP (ISE)) for the identification of high leverage points in linear model", Comput Stat, Vol 31, pp. 859–877

6.Midi, H ., Norazan, M.R ., and Imon ,H.M.(2009) "The Performance of Diagnostic Robust Generalized Potentials for The Identification of Multiple High Leverage Points in Linear Regression", Journal of Applied Statistics, Vol 36:5, PP. 507-520

7.Midi, H ., Sani ,M ., and Arasan , J. (2018) " Robust Heteroscedasticity Consistent Covariance Matrix Estimator based on Robust Mahalanobis Distance and Diagnostic Robust Generalized Potential Weighting Methods in Linear Regression" Journal of Modern Applied Statistical Methods. Vol (17) iss.1

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

9.Rousseeuw, P.J ., Van Zomeren , B.C (1990) " Unmasking multivariate outliers and leverage points ", J Am Stat Assoc, vol 85(411), PP. 633–639

10. Rousseeuw, P.J., Aelst, S.F.(2009) " Minimum volume ellipsoid " John Wiley & Sons, Inc. vol (1) , pp. 71–82

11. White, H. (1980). "A heteroskedasticity-consistent covariance matrix estimator and a direct test for heteroskedasticity". Econometrica, vol 48(4),pp. 817-838.

12. Yan, X., Su, X. G. (2005). "Linear Regression Analysis Theory and Computing" World Scientific Publishing Co. Pte. Ltd .

مقارنة بعض طرائق تقدير معلمات انموذج الانحدار الخطي العام في حالة عدم تجانس تباين الخطأ وظهور القيم الشاذة المتطرفة (HLP_S)

قاسم محمد صاحب
كلية الادارة والاقتصاد ، جامعة بغداد
qasimalsaheb@gmail.com

ا.د. سجى محمد حسين
كلية الادارة والاقتصاد ، جامعة بغداد
saja@coadec.uobaghdad.edu.iq

Received: 16/12/2020 **Accepted:31/1/2021** **Published: March/ 2021**

مستخلص البحث

يعتبر الانحدار الخطي احد اهم الادوات الاحصائية التي يمكن من خلاله معرفة العلاقة بين متغير الاستجابة (Response Variable) ومتغير واحد او عدد من المتغيرات التوضيحية (Independed Variable) ، الذي استخدم في مجالات العلوم المختلفة . عدم تجانس التباين تعد احدى مشاكل الانحدار الخطي والتي يؤدي تأثيرها الى استنتاجات غير دقيقة ، كما قد يرافق مشكلة عدم تجانس التباين وجود القيم الشاذة المتطرفة في المتغير التوضيحي (HLP_S) ، وان وجودها في مجموعة البيانات يؤدي الى تقديرات غير واقعية واستدلالا مضللا .

نستعرض في هذا البحث بعض طرائق التقدير الموزونة الحصينة التي تستوعب كل من الطرائق الحصينة والتقليدية في الكشف عن القيم الشاذة المتطرفة (HLP_S) وتحديد الاوزان ،الطرائق تشمل كل من التشخيص الحصين العام الكامن بناءً على اصغر حجم قطع ناقص ((DRGP(MVE))، التشخيص الحصين العام الكامن بناءً على اصغر محدد مصفوفة تباين مشترك ((DRGP(MCD))، التشخيص الحصين العام الكامن بناءً على مساواة مجموعة الفهرسة ((DRGP(ISE)) ، تمت اجراء المقارنة وفق معيار الخطأ المعياري لمصفوفة معلمات انموذج الانحدار الخطي العام المقدرة ($SE(HC_{W4})$ و$SE(HC_{W5})$ ، ولحجوم عينات (n=60,n=100,n=160) وحدة عدم تجانس مختلفة وبنسب تلوث (HLP_S) هي $(\tau = 10\%, \tau = 30\%)$ ، تبين من خلال المقارنة ان مقدرات المربعات الصغرى الموزونة المعتمدة على اوزان طريقة (DRGP(ISE)) تعتبر الافضل في تقدير معلمات انموذج الانحدار الخطي المتعدد لكونها تمتلك اقل قيم الخطأ المعياري للمقدرين (HC_{W4}) (HC_{W5}) بالمقارنة مع بقية الطرائق .

المصطلحات الرئيسة للبحث : التشخيص الحصين العام الكامن ، مصفوفة التباين والتباين المشترك المتسقة الحصينة ، الاغراق ، الاخفاء .