Remittances Review
April, 2024
Volume: 9, No: 2, pp.3599-3615
ISSN: 2059-6588(Print) | ISSN 2059-6596(Online)
Received : 15 february 2024, Accepted: 05 April 2024
DOI: https://doi.org/10.33282/rr.vx9i2.185

STANDARD OPERATING PROCEDURE FOR EFFICIENT MODEL SELECTION THROUGH HYBRID OF STEPWISE AND ROBUST REGRESSION ANALYSIS

Anam Javaid[1,2], Khadija[3] and Tahira Bano Qasim[4], Shahbaz Nawaz[5]

[1]*School of Mathematical Sciences, Universiti Sains Malaysia*
11800 USM, Penang, Malaysia
[2]*Assistant Professor;Department of statistics, The Women University Multan, Pakistan*
[3]*M.Phil Student, Department of Statistics, The Women University Multan, Pakistan*
[4]*Associate Professor;Department of statistics, The Women University Multan, Pakistan*
[5]*School of Quantitative Sciences, University Utara Maalaysia, 06010 Kedah, Malaysia*

Corresponding Author Email: *tahirabanoQasim@yahoo.com*
anamjavaid7860@yahoo.com

ABSTRACT

Wheat is considered as an important production in agriculture. Because it is included among the basic necessitates for the human livings. Therefore, the production of wheat is important in various field. Especially, in Pakistan there is a need to increase the production of wheat because the population is increasing. There are various factors effecting on the wheat production. The need is to focus on the factors related to wheat. The current study focused on the extraction of factors related to the wheat production. Dataset is used from Statistical Bureau of Pakistan consisting of 466 observations in the analysis. In which 80% dataset is taken as the train dataset while the 20% is kept as a test dataset. 46 predictors related to the yield of wheat is observed with the dependent variable as yield of wheat. Econometric issues such as multicollinearity and outliers are observed in the dataset. For this purpose, hybrid model of robust estimators and forward stepwise regression is used. Among the robust estimators, huber M, hampel M and bisquare M is chosen for the comparison purpose. The results showed that the hybrid model of forward stepwise and hampel M estimators provide the efficient results in term of minimum mean square error (MSE) and mean absolute percentage error (MAPE).

Introduction

Agriculture is considered as the most important sector in Pakistan as mostly of the revenue is gained from this sector in Pakistani economy (Azam and Shafique; 2017). Among the agriculture field, wheat is considered as the main crop in production of

Pakistan (Sher and Ahmad; 2008). The sowing time period if wheat is from October to December and harvesting time period is from March to May in Pakistan (Haider et al., 2019). For the wheat production, various factors has their own importance such as fertilization dose, land type, temperature, environment and such other factors (Haider et al., 2019). As the population of the Pakistan is increasing, so there is a need to focus on the factors related to the yield of wheat production (Iqbal et al., 2015). Such that the production of the wheat can be increased to meet the population requirement at least within the country (Iqbal et al., 2015). For the factors identification regarding the wheat production, various researchers worked with the statistical techniques to get the actual fact and figures (Lobell et al., 2005). Because the forecasting is only possible with the updated statistical techniques (Rao et al., 2016).One of the interesting technique used for the forecasting purpose is the regression analysis (Javaid et al., 2020). Among the regression analysis, Ordinary Least Square (OLS) is one of the simplest technique used for the analysis of factors (Javaid et al., 2019).But it provide inconsistent results in case of its assumption violation (Gujrati; 2022). In such a situations, various other regression techniques are available for the analysis purpose (Karkacier et al., 2006). Among one of the useful technique is the robust regression analysis (Javaid et al., 2021). Robust analysis provide the efficient result even in the case of assumption violation for OLS (Javaid et al., 2019). Also the efficient and consistent results can be obtained in case of outliers are present in the dataset (Rousseeuw and Leroy; 2005). Robust regression contains different kind of estimators such as M, MM, LTS, R estimators (Li, 1985). Among them, M estimators are still have its importance due to providing the efficient estimates as compared to other robust estimators (Bellec and Shen, 2022). The researchers are working with the hybrid techniques to get the advantage of various techniques in a single analysis (Javaid et al., 2020). In the current study, hybrid model selection is done with the help of robust and stepwise estimators. Because of the quality of stepwise estimators to deal with in case of more variables are involved in the dataset (Johnsson; 1992). The contribution of the current study is to develop a Standard Operating Procedure (SOP) to obtain the efficient model through the hybrid of stepwise and robust estimators. As no research has done yet with the hybrid of stepwise and robust estimators in the agricultural field in Pakistan.

METHODOLOGY

The study used the hybrid of stepwise and robust estimators. Among the stepwise regression, forward estimators are chosen for the analysis purpose due to the large number of variables involvement in the dataset. While, for the robust estimators, M estimators are selected to deal with the problem of outliers in the dataset. The details of the methodology used in this research are discussed as follows.

Remittances Review
April, 2024
Volume: 9, No: 2, pp.3599-3615
ISSN: 2059-6588(Print) | ISSN 2059-6596(Online)

Ordinary Least Square

OLS is the simple kind of regression analysis used for the identification of significant variables (Gujrati; 2022). For the simple linear regression estimators, the model can be defined as in equation (1) for the estimation of $\beta_1 and\ \beta_2$:

$$Y_i = n\beta_1 + \beta_2 \Sigma X_i$$

(1)

The normal equations for the estimation purpose are defined as in equation (2) and (3)

$$\sum Y_i X_i = \beta_1 \sum X_i + \beta_2 \sum X_i^2 \quad (2)$$

$$\beta_2 = \frac{n\sum X_i Y_i - \sum X_i Y_i}{n\sum X_i^2 - (\sum X_i)^2} = \frac{\sum x_i y_i}{\sum (x_i)^2}$$

(3)

where $\underline{X}$ and $\underline{Y}$ are the sample means of X and Y and where we define $x_i = (X_i - \underline{X})$ and $y_i = (Y_i - \underline{Y})$. So

$$\beta_1 = \frac{(\sum X_i)^2 \sum Y_i - \sum X_i \sum Y_i}{n\sum X_i^2 - (\sum X_i)^2} = \underline{Y} - \beta_2 \underline{X}$$

(4)

The ultimate step in equation (4) can be obtained direct from (1) with the aid of easy algebraic manipulations. Incidentally, notice that, through utilizing easy algebraic identities, formula (three) for estimating β_2 may be rather expressed as

$$\beta_2 = \frac{\sum x_i y_i}{\sum (x_i)^2} = \frac{\sum X_i Y_i}{\sum X_i^2 - n(\underline{X})^2}$$

The estimators received formerly are called the least square's estimators, for they are derived from the least square's precept.

STEPWISE ESTIMATORS

Stepwise regression is a famous facts-mining device that makes use of statistical significance to select the explanatory variables to be used in a multiple-regression version.

Efroymson (1960) proposed selecting the explanatory variables for a multiple regression version from a set of candidate variables through a chain of automated steps. At every step, the candidate variables are evaluated, one after the other, normally using the t records for the coefficients of the variables being taken into consideration (Lim et al., 2020).

Remittances Review
April, 2024
Volume: 9, No: 2, pp.3599-3615
ISSN: 2059-6588(Print) | ISSN 2059-6596(Online)

ROBUST M ESESTIMATORS

Draper and Smith (1998) defined the procedure for maximum likelihood type estimates (M estimates). Instead of minimizing the sum of squared mistakes, the M-estimates seeks to minimize a function of the errors as in equation (5)

$$min \sum_{i=0}^{n} \rho\left(\frac{e_i}{S}\right) = min \sum_{i=o}^{n} \rho\left(\frac{y_i - X_i'\widehat{\beta}}{S}\right) \quad (5)$$

where S is a scale estimate derived from a linear combination of residuals. The function calculates how much each residual contributes to the goal function. A reasonable ρ have following properties;

$$\rho(e) \geq 0, \rho(e) = \rho(-e), \text{and } \rho(e_i) \geq \rho(e_i') \text{ for } |e_i| \geq |e_i'|.$$

For example, $\rho(e_i) = e_i^2$ in the least squares estimate. Taking partial derivatives with respect to and putting them equal to 0 yields the system of normal equations needed to solve this minimization issue as in equation (6)

$$\sum_{i=o}^{n} \boldsymbol{\psi}\left(\frac{\boldsymbol{Y_i - X_i \beta_i}}{\boldsymbol{S}}\right) \boldsymbol{X_i} = \mathbf{0} \quad (6)$$

Where the derivative of the function is chosen depending on the decision on how much weight to give outliers. Large outliers are not given as much weight by a monotone function as they are by least squares (e.g., a 10 outlier is given the same weight a 3 outlier). A recursive function raises the weight assigned to an outlier until it reaches a certain distance, then drops the weight to 0 as the outlying distance grows. The two approaches for solving the M-estimate nonlinear normal equations are Newton-Raphson and Iteratively Reweighted Least Squares (IRLS). The normal equation is expressed by IRLS as,

$$X'WX\beta = \beta'WY$$

Where W is an n×n diagonal matrix of weights in equation (7)

$$W_i = \frac{\boldsymbol{\psi}\left(\frac{\boldsymbol{Y_i - X_i'\widehat{\beta}_0}}{\boldsymbol{S}}\right)}{\left(\frac{\boldsymbol{Y_i - X_i'\widehat{\beta}_0}}{\boldsymbol{S}}\right)} \quad (7)$$

OLS is commonly used to produce the initial vector of parameter estimations $\hat{\beta}_0$.
These parameters estimations are updated by IRLS.

$$\beta_1 = (X'WX)^{-1}X'WY$$

The weights on the other hand are dependent on the residuals which are independent on the weights. Iteratively re-weighted least squares is an iterative solution. Among the M estimators, Huber M, Hampel M and Tukey Bisquare M estimators are used in this research. The weights assigned to each method is discussed as follows (Stuart; 2011).

Huber's M Estimator

$$p(u) = \begin{cases} u^2 & if|u| < c \\ |2u|c - c^2 & if|u| \geq c \end{cases}$$

$$\psi(u) = \begin{cases} u & if|u| < c \\ c\sin(u) & if|u| \geq c \end{cases}$$

Remittances Review
April, 2024
Volume: 9, No: 2, pp.3599-3615
ISSN: 2059-6588(Print) | ISSN 2059-6596(Online)

$$w(u) = \begin{cases} 1 & if|u| < c \\ c/|u| & if|u| \geq c \end{cases}$$

c=1.345

Hampel M Estimator

$$\boldsymbol{\rho}(\boldsymbol{v}) = \begin{cases} v & if\ |v| < a \\ a|v| - \frac{1}{2}a^2 & if a \leq |v| < b \\ a\frac{c|v| - \frac{1}{2}v^2}{c-b} - \frac{7a^2}{6} & if\ b \leq |v| \leq c \end{cases}$$

$$\psi(v) = \begin{cases} v & if\ |v| < a \\ a\sin v & if a \leq |v| < b \\ a\frac{c\sin v - v}{c-b} & if\ b \leq |v| \leq c \end{cases}$$

$$w(v) = \begin{cases} 1 & if\ |v| < a \\ \frac{a}{|v|} & if a \leq |v| < b \\ o & otherwise \end{cases}$$

Tukey's Biweight M Estimator

$$p(u) = \begin{cases} \frac{c^2}{3}\left\{1 - \left[1 - \left(\frac{u}{c}\right)^2\right]^3\right\} & if|u| < c \\ 2c & if|u| \geq c \end{cases}$$

$$\psi(u) = \begin{cases} u\left[1 - \left(\frac{u}{c}\right)^2\right]^2 & if|u| < c \\ 0 & if|u| \geq c \end{cases}$$

$$w(u) = \begin{cases} \left[1 - \left(\frac{u}{c}\right)^2\right]^2 & if|u| < c \\ 0 & if|u| \geq c \end{cases}$$

c = 4.685

Each of the weighting method has its own properties in providing the efficient results. All three methods will be compared in this research and final model will be chosen with the most efficient results in the analysis.

RESULTS AND DISCUSSION

Data Collection and Procedure

The dataset used for the analysis purpose is taken from Statistical Bureau of Pakistan. Total of 466 observations are taken in the analysis. In which, 80% dataset consisting of 372 observations are used as a train dataset. While 20% consisting of 94 observations are kept as a test dataset for forecasting purpose. The 46 predictors related to the yield is taken for the analysis purpose. The codes are given to all the included variables in the analysis. The list of the codes with their respective names are mentioned in Table 1.

Table 1: *Variables codes and description*

Serial No	**Variable Name**	**Variable Coding**
1	Year	YEAR
2	Division	DIV
3	District	DIST
4	Tehsil	TEHSIL
5	Markaz	MARKAZ
6	Union Council	UC
7	Village	VILL
8	HBNO	HBN
9	Plotno	PN
10	Field Kanal	FK
11	Field Marla	FM
12	Net Land	NL
13	Gr Name	GRN
14	Gr Phone	GRP
15	Latitude	LAT
16	Longitude	LONG
17	Yield	YIELD
18	Population of Plants	POP
19	Cut Date	CD
20	Wheat Varity	WV
21	Seed Source	SS
22	Seed Type	ST
23	Qty Seed (quantity of seed)	QTS
24	Sow Date	SD
25	Sow Mode	SM
26	Gobber	GOB
27	Urea	UREA
28	DAP	DAP
29	Ofert Name	OTN
30	Ofert Qty	OQT
31	Soil Type	ST
32	Irrigation	IRR
33	No Water W (tube well)	NWW
34	No Water Canal (canal)	NWC
35	Start Machine	STM

36	Sow Machine	SME
37	Cut Machine	CM
38	Wtresidual	WTR
39	Last Crop	LC
40	Seed Treat	ST
41	Atak Animal	AA
42	Atak Pes	AP
43	Atak Weed	AW
44	Spray Pest No	SPN
45	Spray Weed No	SWN
46	Harvest Price	HP

Phase I

Due to the large number of predictors, forward stepwise regression analysis is used for the extraction of significant predictors among all in the model through SPSS software. As a result, 6 variables are extracted including POP, UREA, FM, TEHSIL, WTR and SPN as mentioned in Table 2.

Table 2: *List of the extracted variables by using the forward stepwise regression analysis*

Variable/ Step	**1**	**2**	**3**	**4**	**5**	**6**
WTR **p-value**	0.148 0.000	0.115 0.000	0.121 0.000	0.155 0.000	0.137 0.000	0.141 0.000
POP **p-value**		0.001 0.000	0.001 0.000	0.001 0.000	0.001 0.000	0.001 0.000
TEHSIL **p-value**			0.050 0.007	0.087 0.000	0.089 0.000	0.087 0.000
SPN **p-value**				0.146 0.000	0.162 0.000	0.164 0.000
UREA **p-value**					0.002 0.002	0.002 0.002
FM **p-value**						-0.004 0.037

The results in Table 2 show that all the extracted variables are highly significant with *p*-value less than 0.05 and even 0.01. Thus the results are highly significant at 5% and even at 1% level of significance.

Phase II

Multicollinearity is tested among the selected predictors achieved by stepwise regression analysis. Correlation matrix is analyzed for the purpose to check multicollinearity among predictors in the model as in Table 3.

Table 3: *Multicollinearity analysis among selected predictor*

	X1	**X2**	**X3**	**X4**	**X5**	**X6**
X1	1.000	-0.056	0.0526	-0.026	0.099	0.033
X2		1.000	0.357	0.154	0.395	0.053
X3			1.000	-0.160	0.351	0.167
X4				1.000	0.267	-0.262
X5					1.000	-0.263
X6						1.000

As represented in Table 3, all the correlation coefficients are less than 0.95, thus no multicollinearity is diagnosis between the selected predictors (Javaid et al., 2020). After multicollinearity test, outlier diagnosis measure is checked through the boxplot analysis. Outliers are found in some predictors such as YIELD, POP and WTR.

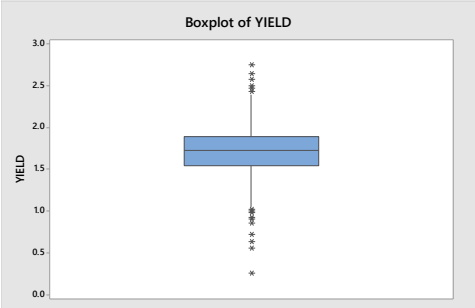

Figure 1: Boxplot of Yield

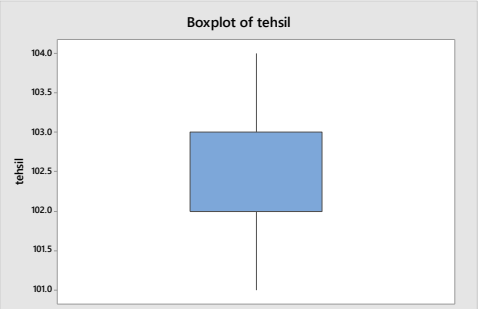

Figure 2: Boxplot of Tehsil

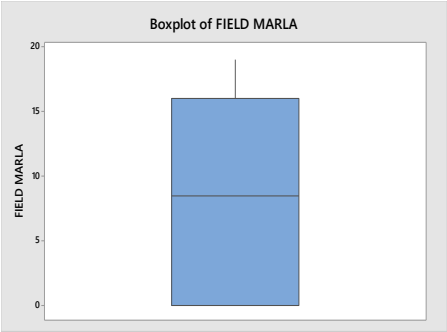

Figure 3: Boxplot of Field Marla

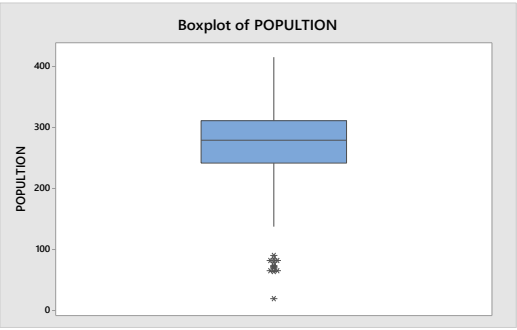

Figure 4: Boxplot of Population

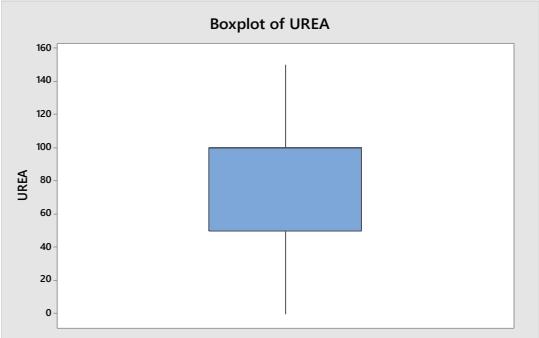

Figure 5: Boxplot of Urea

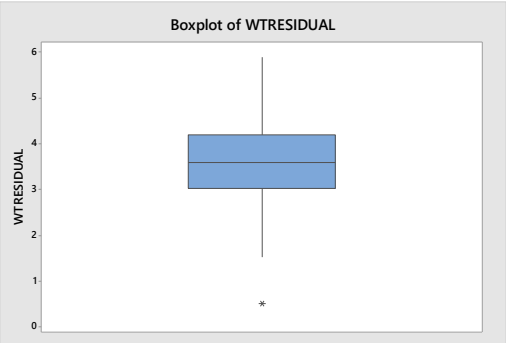

Figure 6: Boxplot of Wtresidual

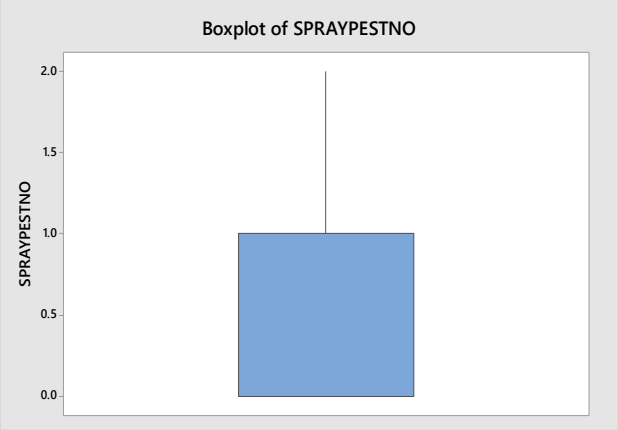

Figure 7: Box plot of Spraypestno

In Figures 1-7, the results can be observed for outliers in various predictors. The dots outside the boxplot represent the outliers in that variable.

Phase III

Due to the presence of outliers in the dataset, robust regression analysis is carried out in Phase III for the purpose of getting efficient results. Because the OLS cannot perform in case of outliers in the dataset (Gujrat; 2022). For the robust regression analysis purpose, Huber M, Hampel M and Tukey Bisquare M is used through R software. The results for Huber M are observed in Table 4.

Table 4: *Results for the method of Huber-M estimator*

Factors	**Coefficients**	***p*-value**	**Significance**
Intercept	-7.399	$3.7167e^{-05}$	Significant
TEHSIL	0.079	$4.8056e^{-06}$	Significant
FM	-0.004	$3.5246e^{-02}$	Significant
POP	0.001	$8.2323e^{-05}$	Significant

UREA	0.002	$2.5338e^{-03}$	Significant
WTR	0.165	$9.6036e^{-19}$	Significant
SPN	0.183	$1.1482e^{-08}$	Significant

From Table 4, the final selected model by using the Huber-M estimator regression method can be defined as in equation (8)

$$\hat{Y} = -7.3993 + 0.0789TEHSIL - 0.0036FM + 0.0009POP + 0.0017UREA + 0.1649WTR + 0.1828\,SPN \quad (8)$$

Output of the regression shows that the TEHSIL, FM, POP, UREA, WTR and SPN predictor variables are statistically significant because *p*-value of the predictor variables are less than 0.05. There will be 0.0789 unit increase in the yield of wheat by a unit increase in Tehsil. The next significant factor is FM, which shows that 0.0036unit decrease in the yield of wheat by a unit decrease in FM. By the empirical results, the other significant factor is population. This shows the 0.0009-unit increase in the yield of wheat by a unit increase in population. That is if one plant will increase, there will be 0.0009 KG unit increase in the yield of wheat production. Through analysis results, the further significant factor is UREA. That shows the 0.0017 KG unit increase in the yield of wheat by a unit increase in UREA. There will be 0.1649 KG unit increase in the yield of wheat by a unit increase in WTR. The next significant factor is SPN, by the analysis results. That shows the 0.1828unit increase in the yield of wheat by a unit increase in SPN.

Second analysis is carried out by using the Hampel M estimator in R software. The results are obtained in Table 5.

Table 5: *Result for the method of Hampel-M estimator*

Factors	**Coefficients**	***p*-value**	**Significance**
Intercept	-7.786	$6.3892e^{-05}$	Significant
TEHSIL	0.082	$1.0940e^{-05}$	Significant
FM	-0.004	$3.5131e^{-02}$	Significant
POP	0.001	$1.4765e^{-05}$	Significant
UREA	0.002	$3.1778e^{-03}$	Significant
WTR	0.158	$2.8986e^{-15}$	Significant

SPN	0.181	$1.7406e^{-07}$	Significant

From Table 5, the final selected model by using the Hampel-M estimator regression method can be defined in equation (9)

$$\hat{Y} = -7.7857 + 0.0823TEHSIL - 0.0039FM + 0.0011POP + 0.0018UREA + 0.1582WTR + 0.1812\,SPN \quad (9)$$

Output of the regression shows that the TEHSIL, FM, POP, UREA, WTR and SPN predictor variables are statistically significant as *p*-value of the predictor variables are less than 0.05. This shows that 0.0823 unit increase will occur in the yield of wheat by a unit increase in Tehsil. The next significant factor is FM, which shows that 0.0039unit decrease in the yield of wheat by a unit decrease in FM. By the analysis results, the other significant factor is population. That shows the 0.0011-unit increase in the yield of wheat by a unit increase in population. That is if one plant will increase, there will be 0.0011 KG unit increase in the yield of wheat production. Through analysis results, the further significant factor is UREA. That shows the 0.0018 KG unit increase in the yield of wheat by a unit increase in UREA. There will be 0.1582 KG unit increase in the yield of wheat by a unit increase in WTR. The next significant factor is SPN, by the analysis results. That shows the 0.1812unit increase in the yield of wheat by a unit increase in SPN.

In robust regression analysis, Tuckey bisquare is carried out in R software. The results are obtained in term of Table 6.

Table 6: *Result for the method of Tukey Bisquare-M estimator*

Factors	**Coefficients**	***p*-value**	**Significance**
Intercept	-7.362	$8.1970e^{-05}$	Significant
TEHSIL	0.079	$1.2066e^{-05}$	Significant
FM	-0.003	$5.3934e^{-02}$	Non-Significant
POP	0.001	$1.1819e^{-03}$	Significant
UREA	0.002	$9.0869e^{-03}$	Significant
WTR	0.174	$3.6497e^{-19}$	Significant
SPN	0.193	$8.1547e^{-09}$	Significant

From Table 6, the final selected model by using the Bisquare-M estimator regression method can be defined in equation (10)

$$\hat{Y} = -7.3624 + 0.0787TEHSIL + 0.0008POP + 0.0015UREA + 0.1744WTR + 0.1927SPN \quad (10)$$

Output of the regression shows that the TEHSIL, POP, UREA, WTR and SPN predictor variables are statistically significant because *p*-value of the predictor variables are less than 0.05. On the contrary, FM is not statistically significant because the *p*-value is greater than 0.05 (Lodhi et al., 2020). There will be 0.078unit increase in the yield of wheat by a unit increase in Tehsil. By the analysis results, the other significant factor is population. That shows the 0.0008-unit increase in the yield of wheat by a unit increase in population. That is if one plant will increase, there will be 0.0008 KG unit increase in the yield of wheat production. There will be 0.0041unit decrease in the wheat yield. Through analysis results, the further significant factor is UREA. That shows the 0.0015 KG unit increase in the yield of wheat by a unit increase in UREA. There will be 0.1744 KG unit increase in the yield of wheat by a unit increase in WTR. The next significant factor is SPN, by the analysis results. That shows the 0.1927unit increase in the yield of wheat by a unit increase in SPN.

Comparison with OLS

For the comparison purpose, OLS is used on the all the selected variables through forward regression analysis. The results are observed in Table 7.

Table 7: *Result for the method ordinary least square*

Factors	**Coefficient**	***p*-value**	**Significance**
Intercept	-8.275	$7.66e^{-05}$	Significant
TEHSIL	0.087	$1.59e^{-05}$	Significant
FM	-0.004	$3.70e^{-01}$	Non-Significant
POP	0.001	$3.40e^{-07}$	Significant
UREA	0.002	$2.20e^{-02}$	Significant
WTR	0.141	$3.21e^{-11}$	Significant
SPN	0.164	$1.01e^{-05}$	Significant

From Table 7, the final selected model by using the OLS regression method can be defined in equation (11)

$$\hat{Y} = -8.2751 + 0.0868TEHSIL + 0.0013POP + 0.0020UREA + 0.1411WTR + 0.1636\,SPN \quad (11)$$

Output of the regression shows that the TEHSIL, POP, UREA, WTR and SPN predictor variables are statistically significant because *p*-value of the predictor variables are less than 0.05 (Javed et al., 2022). On the contrary, FM is not statistically significant because the *p*-value is greater than 0.05. There will be 0.0868unit increase in the yield of wheat by a unit increase in Tehsil. By the analysis results, the other significant factor is population. That shows the 0.0013-unit increase in the yield of wheat by a unit increase in population. That is if one plant will increase, there will be 0.0013 KG unit increase in the yield of wheat production. There will be 0.0041unit decrease in the wheat yield by a unit decrease in FM. Through analysis results, the further significant factor is UREA. That shows the 0.0020 KG unit increase in the yield of wheat by a unit increase in urea. There will be 0.1411 KG unit increase in the yield of wheat by a unit increase in WTR. The next significant factor is SPN, by the analysis results. That shows the 0.1636-unit increase in the yield of wheat by a unit increase in SPN. Overall comparison is done in Table 8.

Table 8: *Comparison between the results of M-Estimators and OLS*

Significance Variable	**Coefficients**			
	OLS	**Robust Regression** **Huber**	**Hampel**	**Bisqaure**
TEHSIL	0.0868	0.0790	0.0823	0.0787
FM	--------	-0.0036	-0.0039	-------
POP	0.0013	0.0009	0.0010	0.0007
UREA	0.0020	0.0017	0.0018	0.0015
WTR	0.1411	0.1649	0.1582	0.1743
SPN	0.1636	0.1828	0.1811	0.1927
	P-values			
TEHSIL	$1.59e^{-05}$	$4.8056e^{-06}$	$1.0940e^{-05}$	$1.2066e^{-05}$
FM	$3.70e^{-01}$	$3.5246e^{-02}$	$3.5131e^{-02}$	$5.3934e^{-02}$
POP	$3.40e^{-07}$	$8.2323e^{-05}$	$1.4765e^{-05}$	$1.1819e^{-03}$
UREA	$2.20e^{-02}$	$2.5338e^{--03}$	$3.1778e^{-03}$	$9.0869e^{-03}$
WTR	$3.21e^{-11}$	$9.6036e^{-19}$	$2.8986e^{-15}$	$3.6497e^{-19}$
SPN	$1.01e^{-05}$	$1.1482e^{-08}$	$1.7406e^{-07}$	$8.1547e^{-09}$

The non-significant factors are already excluded from the final selected model on the basis of *p*-value. OLS excluded one non-significant variable FM from the model. Huber

and Hampel M-estimator included all the significant variables in the model and Bisquare excluded one non-significant variable FM to the model. At 5% level of significance, TEHSIL, FM, POP, UREA, WTR and SPN are significant in the OLS model. In Huber M-estimator six variables are significant, at 5% level of significance. Six variables are significant in Hampel M-estimator at five percent level of significance. In Tukey Bisquare
estimator five variables are significant at 5% level of significance. In OLS, the coefficients are different as compared to the robust regression analysis. While all the tree robust estimators are providing almost the same coefficients in the regression analysis.

Phase IV

For the efficient model selection, MSE and MAPE are observed for each selected models through different kind of estimators. The results can be observed in Table 9.

Table 9: *Results of MSE and MAPE for different estimators*

	MSE	**MAPE**
OLS	0.0822	14.64
HUBER	0.0001	14.63
HAMPEL	0.0001	14.23
BISQUARE	0.0002	20.60

Table 8 shows that the MSE is higher in OLS as compared to the robust estimators. MAPE is lower for Hampel M estimators as compared to the other estimators (Javaid et al., 2020). Thus, the efficient model is selected for the "Yield of wheat" through the Hampel M estimators with the significant factors as TEHSIL, YIELD, POP, UREA, WTR, FM and SPN.

The standardized residual graph is observed for the efficient model selection as in Figure 8

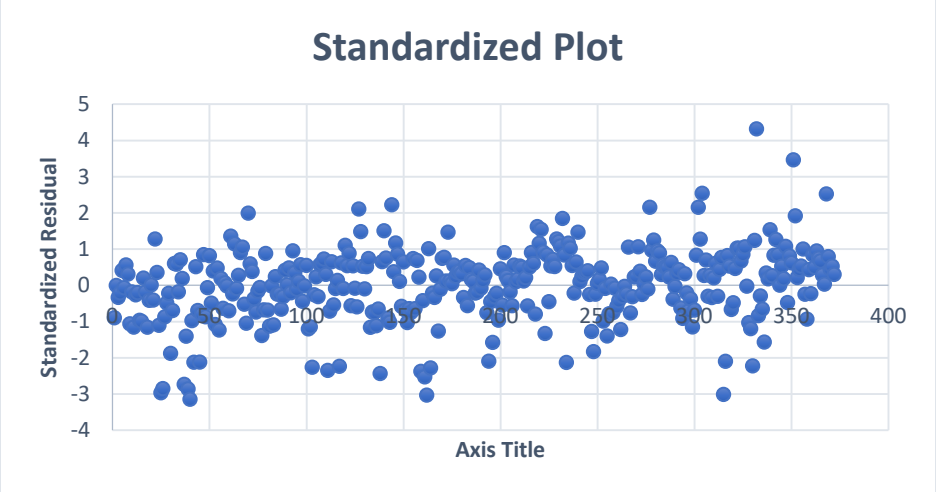

Figure 8: Standardized Residual Plot

The graph shows that there are some outlies found in 3 sigma limits thus the final efficient selected model through Hampel M estimator can be used for forecasting the "Yield of Wheat" through the YIELD, UREA, TEHSIL, POP, WTR, FM and SPN as independent variables.

CONCLUSION

From the obtained results through *Phase I* to *Phase IV*, efficient model selection is observed through the Hampel M estimators. As the MSE and MAPE are in favor of the Hampel M as compared to the other estimators used in this research project. Thus the forecasting for the "Yield" can be obtained efficiently by just more focusing on the YIELD, UREA, POP, FM, TEHSIL, WTR and SPN variables.

Robust M-estimators have been applied in the data set of wheat, which have 45 independent variables and one dependent variable. After stepwise forward regression analysis six variables was selected. Correlation matrix and box plots was used to analyzed the multicollinearity and outliers respectively. No multicollinearity was deducted in the analysis while the boxplot reveals that there are outliers present in the dataset. Due to the presence of outliers in the dataset, robust regression was used for the purpose of analysis. The approaches proposed in this paper are OLS, Huber, Hampel and Bisquares. The comparison results showed that OLS results are not good, so we can not rely on it. The findings showed that Hample-M estimators was best model for forecasting rather than others. Because its MSE and MAPE was less than others.

References

Azam, A., & Shafique, M. (2017). Agriculture in Pakistan and its Impact on Economy. *A Review. Inter. J. Adv. Sci. Technol*, *103*, 47-60.

Bellec, P. C., & Shen, Y. (2022, June). Derivatives and residual distribution of regularized M-estimators with application to adaptive tuning. In *Conference on Learning Theory* (pp. 1912-1947). PMLR.

Draper, N. R., & Smith, H. (1998). *Applied regression analysis* (Vol. 326). John Wiley & Sons.

Efroymson MA. Multiple regression analysis. In: Ralston A, Wilf HS, editors. Mathematical methods for digital computers. New York: Wiley; 1960.

Gujarathi, D. M. (2022). *Gujarati: Basic Econometrics*. McGraw-hill.

Haider, S. A., Naqvi, S. R., Akram, T., Umar, G. A., Shahzad, A., Sial, M. R., Khaliq, S., & Kamran, M. (2019). LSTM neural network-based forecasting model for wheat production in Pakistan. *Agronomy*, *9*(2), 72.

https://www.pbs.gov.pk

https://www.finance.gov.pk

Iqbal, M. A., Iqbal, A., Afzal, S., Akbar, N., Abbas, R. N., & Khan, H. Z. (2015). In Pakistan, agricultural mechanization status and future prospects. *American-Eurasian Journal of Agricultural & Environmental Sciences*, *15*(1), 122-128.

Javaid, A., Ismail, M. T., & Ali, M. K. M. (2019, December). Model selection for collector efficiency of seaweed drier by using LASSO and multiple regression analysis using 8sc. In *AIP Conference Proceedings* (Vol. 2184, No. 1, p. 050032). AIP Publishing LLC.

Javaid, A., Ismail, M., & Ali, M. K. M. (2020). Efficient Model Selection of Collector Efficiency in Solar Dryer using Hybrid of LASSO and Robust Regression. *Pertanika Journal of Science & Technology*, *28*(1).

Javaid, A., Ismail, M. T., & Ali, M. K. M. (2021). Efficient Model Selection For Moisture Ratio Removal Of Seaweed Using Hybrid Of Sparse And Robust Regression Analysis. *Pakistan Journal of Statistics and Operation Research*, 669-681.

Javaid, A., Ismail, M. T., & Ali, M. K. M. (2020). Comparison of sparse and robust regression techniques in efficient model selection for moisture ratio removal of seaweed using solar drier. *Pertanika Journal of Science and Technology*, *28*(2), 609-625.

Javed, Z., Javaid, A., Javaid, S., & Javed, A. (2022). Analysis of Child Mortality on Geographical Basis Over Different Centuries. *STATISTICS, COMPUTING AND INTERDISCIPLINARY RESEARCH*, *4*(2), 1-7.

Johnsson, T. (1992). A procedure for stepwise regression analysis. *Statistical Papers*, *33*(1), 21-29.

Karkacier, O., Goktolga, Z. G., & Cicek, A. (2006). A regression analysis of the effect of energy use in agriculture. *Energy Policy*, *34*(18), 3796-3800.

Lobell, D. B., Ortiz-Monasterio, J. I., Asner, G. P., Matson, P. A., Naylor, R. L., & Falcon, W. P. (2005). Analysis of wheat yield and climatic trends in Mexico. *Field crops research*, *94*(2-3), 250-256.

Lodhi, I., Nawaz, S., Javaid, A., Javaid, S., & Javaid, A. (2023). Geographically Analysis of Wheat Production on Annual Basis. *STATISTICS, COMPUTING AND INTERDISCIPLINARY RESEARCH*, *5*(1), 29-36.

Li, G. (1985). Robust regression. *Exploring data tables, trends, and shapes*, *281*, U340.

Lim, H. Y., Fam, P. S., Javaid, A., Ali, M., & Khan, M. (2020). Ridge Regression as Efficient Model Selection and Forecasting of Fish Drying Using V-Groove Hybrid Solar Drier. *Pertanika Journal of Science & Technology*, *28*(4).

Rao, A. L., & Ketema, H. (2016). Statistical analysis of factors affecting wheat production a case study at walmara woreda. *International Journal of Engineering and Management Research (IJEMR)*, *6*(5), 43-53.

Rousseeuw, P. J., & Leroy, A. M. (2005). *Robust regression and outlier detection*. John wiley & sons.

Sher, F., & Ahmad, E. (2008). Forecasting wheat production in Pakistan.

Stuart, C. (2011). *Robust Regression.* Durham, England: Durham University.