EXPLORING THE FACTORS RELATED TO THE YIELD OF SUNFLOWER CROP: AN APPLICATION OF ROBUST AND RIDGE REGRESSION ANALYSIS

IQRA GULSHAN [1], DR. ANAM JAVAID[2],ZAINAB JAVED [3], DR. SHAHBAZ NAWAZ[4]

[1.] M.Phil Scholar, Department of Statistics, The Women University Multan

[2.] Assistant Professor; Department of Statistics, The Women University Multan, Pakistan Email: anamjavaid7860@yahoo.com Corresponding Author

[3.] M.Phil Scholar, Department of Statistics, The Women University Multan

[4]. Punjab Bureau of Statistics; Planning and Development Department, Pakistan

Abstract

The first production of sunflower was introduced to Pakistan in 1960's and was one of the country primary oilseed crops. According to Food Agriculture Organization (FAO), 152,675 hectares of sunflower were cultivated in Pakistan in 2014 and a total of 3240 tons seeds were produced. Pakistan is yielding 0.14 million tons production of sunflower only in the world. There are many factors that contribute to increase yield per hectare, but fertilizer management is more important to enhance sunflower crop growth development and achene (one-seeded dry fruit) production. The current study focus on selection of efficient model for sunflower production in Pakistan. For this purpose, sunflower dataset is taken for analysis. The yield is taken as dependent variable and all other factors such as village name, plant population, soil type ,quantity of seed, urea, Dap, No water well / Tube well, usage of machine, weight residual, last crop, seed treated, Attack pest and harvest price are taken as independent variables. Three steps are used to select an efficient model. In the first step, Correlation matrix and box Plots are used to analyze the multicollinearity and outliers respectively. No multicollinearity among predictors is deducted while the boxplot reveals that there are outliers present in the dataset. In the second steps, due to the presence of outliers in the datasets, Robust Regression will be used for the purpose of analysis. Three M-estimators are used of robust regression (Huber M-estimator, Hampel M-estimator and Tukey bisquare M-estimator). Final steps were considered on the basis of efficient model selection by using model selection criteria such as Mean square error (MSE), Mean Absolute Percentage Error (MAPE), Akaike's information criteria (AIC) and Bayesian information criteria (BIC). The efficient model for "yield of sunflower" is selected by Hampel M estimator and is preferred on the basis of minimum value of MSE, MAPE, AIC and BIC.

Keywords: *Robust regression, OLS, Outliers, Multicollinearity, Model Selection.*

INTRODUCTION

Helianthus annuus L, an annual sunflower is the variety to which most people prefer. It is a type of an annual plant with a huge inflorescence (flowering head).The name comes from the flower's appearance which is frequently interpreted to represent the sun by (Khaleghizadeh 2011, *Fabian et al*; January 2014).Sunflower are produced in a developing areas worldwide (about 28 million hectares in plant with a huge inflorescence flowering head). Sunflower production is increasing (28 million hectares in 2021 with a yield of about 50 million tons NSA) and as a result, its importance is increased day by day in human and animal nutrition by (Vldimir Miklic; April 2022).Sunflower on a big scale with two thirds of the output being concentrated in Europe, particularly Ukraine, Russia and the Turkey region of Trakya .South east Africa (South Africa, Tanzania, Uganda and Zambia) and Argentina, China and the United States and in other different Countries. In India, the acreage was significant but it rapidly decreased from 2.35 in 2006 to 0.5 10 year later and 0.28 million hectares in 2019. The top ten countries including Ukraine, the Russian federation, Argentina, China, Romania, Bulgaria, Turkey, Hungary, France and the united states account for 84% of production and 76% of acreage between 2014 and 2018 by (Etienne Pilorge; June 2020).

The first production of sunflower was introduced to Pakistan in 1960's and was one of the country primary oilseed crops. In addition, other oilseed crops grown in Pakistan includes cottonseed, rapeseed/mustard, canola etc. During the 2007-2008 growth period,(it was planted on an area of 457.30 thousand hectare, resulting in of 264 thousand tons of oil and 683 thousand tons of seed) by (Anonymous 2008, Qureshi et al; 2015). According to Food Agriculture Organization (FAO) 152,675 hectares of sunflower were cultivated in Pakistan in 2014 and a total of 3240 tons seeds were produced. Sunflower is an annually cross-pollinated plant by (Hafiz et al; 2021).The sunflower production nitrogen is considered to be primary nutrient than any other nutrient. In common urea, nitrogen fertilizer is used to meet the nitrogen demand of crops. However one fourth of nitrogen is lost in environment in the form of denitrification (loss of nitrogen), nitrate leaching and ammonia volatilization from common urea .That's why farmers have to apply extra nitrogen dose to less the monetary returns by (Sonia et al; 2021).

In Pakistan, The estimate for the production of cooking oil is 0.503 million tons and the use of cooking oil is 2.447 million tons. It spent Rs.155.278 billion on 1.944 million tons of cooking oil during the 2017-2018 annual year due to increase in its demand (GOP 2017-2018). There are several causes of low yields of sunflower crops but the main ones are delayed sowing times, poor planting methods, drought and heat stress by (Ahmad et al; 2020). Compared to other countries, low production of sunflower in Pakistan is due to insufficient and inappropriate fertilizers. There are many factors that contribute to increase yield per hectare, but fertilizer management is more important to enhance sunflower crop growth development and achene (one-seeded dry fruit) production by (Ahmad et al; 2018).

During 1980's, sunflower were introduced by Ghee Corporation of Pakistan (GCP). In Pakistan sunflower crop was took horizontal expansion 150 times more due to increase in area as compared to 1970-71 where 300 times expansion was on vertical aspects by (Tabassum et al., 2020).

Table 1: Production of sunflower from 1970-2019 in Pakistan.

Year	Area (000,ha)	Production (000, tons)	Yield (t/h)
1971-72	1.25	0.87	0.70
1981-82	7.24	5.86	0.81
1991-92	63.33	83.31	1.32
2001-02	63.24	73.96	1.17
2005-06	325.08	348.28	1.07
2006-07	323.07	407.22	1.26
2007-08	397.31	603.89	1.52
2008-09	319.74	420.49	1.32
2009-10	256.12	325.48	1.27
2010-11	300.61	404.39	1.35
2011-12	236.00	282.92	1.20
2017-18	82.20	104.00	1.27

Source: Agriculture Statistics of Pakistan, Economic survey of Pakistan.

Table 1, provided data on sunflower cultivation in Pakistan over the years, including the area in thousands of hectares, production in thousands of metric tons, and the yield in tons per hectare for each year within he specified time range.

(In 2018-19, main regions of sunflower crop were in the province of Punjab (DG Khan, Bahawalpur and Multan). In DG Khan, maximum crop about 36.58 thousand tons of sunflower was sown with

18.33 thousand hector. Bahawalpur gave 7.84 thousand tons crop in the 4.12 thousands hector by (Tabassum et al., 2020).

Table 2: The region of sunflower crops in Punjab provinces (2018-2019)

S.No.	Region	Area (000 h)	Production (000 t)	Yield (t/h)
1	DG Khan	18.33	36.58	2.00
2	Bahawalpur	4.12	7.84	1.90
3	Multan	3.30	6.01	1.82
4	Sargodha	2.90	3.92	1.35
5	Faisalabad	0.41	0.82	2.00
6	Gujranwala	0.38	0.67	1.76
7	Lahore	0.24	0.45	1.88
8	Sahiwal	0.23	0.45	1.96
9	Rawalpindi	0.01	0.02	2.00
10	**Punjab**	**29.92**	**56.78**	**1.90**

Source: Crop Reporting Service, Punjab, Lahore. 2018-19

Table 2, represents the information about different regions in Punjab provinces, including their area in thousands of hectors, production in thousands of metric tons, and yield in tons per hectare.

REGRESSION ANALYSIS

Regression analysis was a simple method which was used for investigating the functional relationship among the variable. Regression analysis is the term used to describe a group of methods used for modeling numeric data (Shi, R., & Conrad, S. A.; 2009). This relationship was expressed in the term of equation or model connecting the dependent variable with predicator's variables.

Let's denoted the response variable (dependent variable) by Y and the set of predictor's variables $X_1, X_2, \ldots, X_P$ where p denotes the number of predictor's variable. The true relationship between can be calculated by regression model as follow

$$Y = f\ (X_1, X_2, \ldots, X_P) + \mathcal{E},$$

Where $\mathcal{E}$ was a random error, its result as failure of the model to fit the exact data. The function f $(X_1, X_2, \ldots, X_P)$ describe the relationship between Y and $X_1, X_2, \ldots, X_{P.}$ With Linear regression model

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_p X_p + \mathcal{E},$$

Where $\beta_0, \beta_1, \ldots, \beta_P$ called the regression parameters or coefficients these are unknown constants estimated from the data. Unknown constants were commonly used notational convention to denote the unknown parameters by Greek letters by (Chatterjee and Hadi: 2006).

THE METHOD OF ORDINARY LEAST SQUARE

The standard form of OLS matrix notation is given below:

$$\begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ \vdots \\ Y_n \end{bmatrix}_{n\times 1} = \begin{bmatrix} 1 & X_{11} & X_{21} & \cdots & X_{k1} \\ 1 & X_{12} & X_{22} & \cdots & X_{k2} \\ \vdots & \vdots & \vdots & & \vdots \\ \vdots & \vdots & \vdots & \cdots & \vdots \\ 1 & X_{1n} & X_{2n} & \cdots & X_{kn} \end{bmatrix}_{n\times k} \begin{bmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \vdots \\ \beta_n \end{bmatrix}_{k\times 1} + \begin{bmatrix} \mathcal{E}_1 \\ \mathcal{E}_2 \\ \vdots \\ \vdots \\ \mathcal{E}_n \end{bmatrix}_{n\times 1}$$

$$Y = X\beta + \mathcal{E}$$

This equation has systematic component (Xβ) and a stochastic component (Ɛ). Through this equation population parameters can be estimated in β vectors. The OLS method is famous and oldest used techniques in statistical strategies. The result of OLS estimators is based upon unbiased, minimum variance, minimal mean square error, performance and BLUE (best linear unbiased estimators). Basically, least square are used to get numerical value of parameters in features this is known as OLS. The equation of OLS is given as below;

$$\hat{\beta}_{\text{OLS}} = (X'X)^{-1}\ X'Y$$

German mathematician Carl Friedrich Gauss in 1809 introduced the method of ordinary least square. From his assumptions, the method of least square is very impressive statistical properties. OLS is considered most powerful and popular method of regression analysis (Gujarati, D. N; 2004).

ROBUST REGRESSION ANALYSIS

Robust regression works on alternative to the least square regression having least restrictive assumptions (Andersen; 2008). It gives best regression coefficient estimates when outliers are there in dataset. Basically there is violation in assumption of normally distributed residuals with least square regression by outliers. Outliers are very difficult to identify because they have small residuals than there should be. Outlier's points are visually detected in various scatter plots when there is use of one or two independent variables (Rousseeuw et al; 2005). An additional independent variables may hide the outliers from scatter plots. Therefore, robust regression weight down the influence of outliers because of iterative procedure that not only identify outliers but minimize the impact on the estimation of coefficient. This weight which is given to every observation in robust regression is controlled by special curve not an influence function. Robust regression has its own residual analysis and down weights of the outliers. There are many types of estimator available in case of robust regression, but the common type is M-estimation where Huber, Hampel and bisquares are mostly used (Javaid et al; 2020).

Table 3: Weight function used for different regression method

	Objective function $\boldsymbol{\rho(u)}$	**Score function**	**Weight function $\boldsymbol{W(u)}$**
a) Huber M a>0	$\begin{cases} u^2 & if\ \lvert u\rvert < 0 \\ \lvert 2u\rvert c - c^2 & if\ \lvert u\rvert \geq 0 \end{cases}$	$\begin{cases} u & if\ \lvert u\rvert < 0 \\ c\ \text{sign}\ (u) & if\ \lvert u\rvert \geq 0 \end{cases}$	$\begin{cases} 1 & if\ \lvert u\rvert < 0 \\ c/\lvert u\rvert & if\ \lvert u\rvert \geq 0 \end{cases}$
b) Hampel M a,b,c> 0	$\begin{cases} u & if\ \lvert u\rvert < \\ a\lvert u\rvert - \frac{1}{2}a^2 & if\ a\ \leq \\ a\frac{c\lvert u\rvert - \frac{1}{2}u^2}{c-b} - \frac{7a^2}{6} & if\ b \leq \end{cases}$	$\begin{cases} u & if\ \lvert u\rvert < \\ a\ \text{sign}\ u & if\ a \leq \lvert \\ a\frac{c\ sign\ u - u}{c-b} & if\ b \leq \end{cases}$	$\begin{cases} 1 & if\ \lvert u\rvert < \\ \frac{a}{\lvert u\rvert} & if\ a \leq \lvert \\ 0 & otherwise \end{cases}$
c) Tukey Bisquare M a>0	$\begin{cases} \frac{c^2}{3}\left\{1 - \left[1 - \left(\frac{u}{c}\right)^2\right]^3\right\} & if\ \lvert u\rvert < \\ 2c & if\ \lvert u\rvert \geq \end{cases}$	$\begin{cases} u\left[1 - \left(\frac{u}{c}\right)^2\right]^2 & if\ \lvert u\rvert \\ 0 & if\ \lvert u\rvert \end{cases}$	$\begin{cases} \left[\ 1 - \left(\frac{u}{c}\right)^2\right]^2 & if\ \lvert u\rvert \\ 0 & if\ \lvert u\rvert \end{cases}$

RIDGE REGRESSION

Ridge regression was first introduced by Hoerl and Kennard in 1970, to deals with the issues of multicollinearity in the dataset by (Dorugade, A. V. 2014). Basically ridge regression not only reduces the errors but also provides more appropriate estimation for multicollinearity issue than OLS (Gujarati, D.N; 2004).

Consider the regression model

$$Y = X\beta + \varepsilon \qquad (1)$$

Where y is n × 1 vector of dependent variable and ß is p × 1vector of an unknown parameter and Ɛ is n × 1 vector of random error of zero mean and constant variance.

$$\hat{\beta}_{OLS} = (X'X)^{-1}\ X'Y \qquad (2)$$

In equation 2, if the regressors are dependent, matrix $X'X$ becomes ill conditioned. So, Hoerl and Kennard suggest a ridge estimator as in equation 3,

$$\hat{\beta}_{R} = \ (X'X + kI)^{-1}\, X'Y \qquad (3)$$

Where k > 0 is a constant and known as Ridge parameter, shrinkage parameter and biasing parameter.

To form a new matrix $(X'X + \lambda I)$ ridge regression adds a ridge parameter (λ). It is known as ridge regression because the diagonal of correlation matrix must be describe as a ridge. The benefits of ridge regression is that coefficients are shrink and there is less model complexity. It does not require unbiased estimators that's why there is addition of bias estimators to reduce standard error by (Lim, H. Y et al; 2020).

FLOW CHART OF THE METHODOLOGY

For the analysis purpose, various econometrical issues are addressed in the dataset, then the statistical analysis is carried out. The flow chart of the study can be observed in figure 1 as follows,

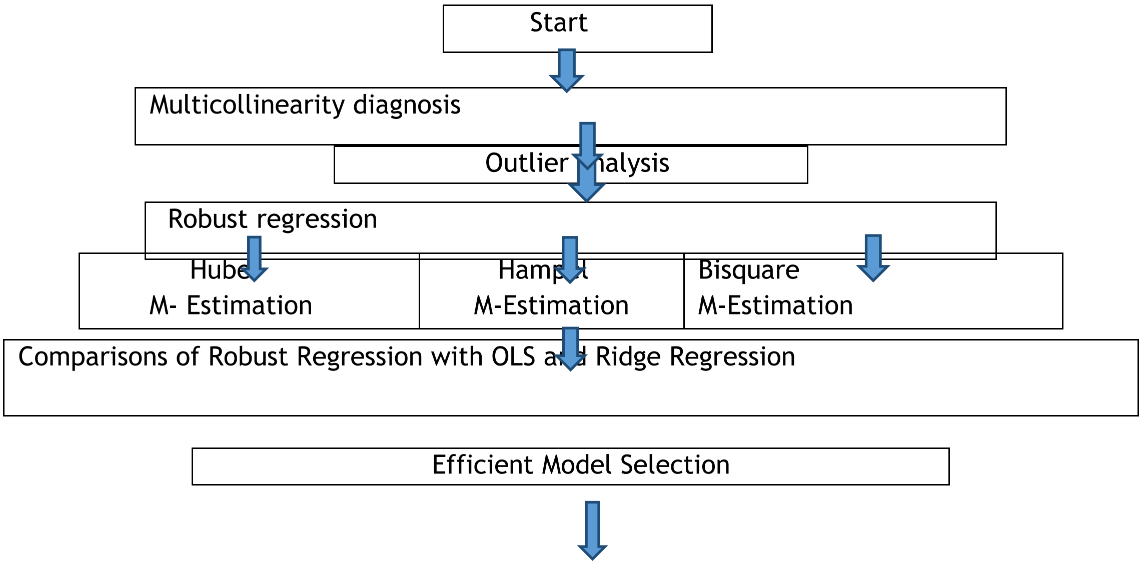
Start
Multicollinearity diagnosis
Outlier Analysis
Robust regression
Huber
M- Estimation
Hampel
M-Estimation
Bisquare
M-Estimation
Comparisons of Robust Regression with OLS and Ridge Regression
Efficient Model Selection

Figure 1: Flow Chart of the Methodology

MATERIALS AND METHODS

The dataset of variables used in this study is taken from Statistical Bureau of Pakistan. Total of 141 observations are used for the purpose of analysis. There are 14 variables related to the yield of sunflower is analyzed in the dataset. Yield is taken as dependent variable and all other factors are such as village name, population, soil type ,quantity of seed, urea, Dap, No water well / Tube well, start machine, weight residual, last crop, seed treated, Attack pest and harvest price are taken as independent variables. There are 141 observations taken in analysis. No missing observation is found in analysis. 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 4 as follow.

Table 4: variables codes and description

Serial No	**Variable Name**	**Variable coding**
1	Yield (YED)	Y
2	Village Name (VN)	X_1
3	Plant Population (PP)	X_2
4	Quantity of Seed (QTS)	X_3
5	Urea (UR)	X_4
6	Dap (DAP)	X_5

7	Soil type (ST)	X_6
8	No- of water well / Tube well (NW)	X_7
9	Usage of Machine (UOM)	X_8
10	Weight Residual (WR)	X_9
11	Last Crop (LC)	X_{10}
12	Seed Treated (ST)	X_{11}
13	Attack Pest (AKP)	X_{12}
14	Harvest Price (HP)	X_{13}

Table 4, represents the factors related to the yield of sunflower. The unit of the yield is taken as in kilogram (KG) while the unit of plant population also in kg. The units of quantity of seed, urea and dap are represent in the term of kg/acre. On the other hand Soil type (silt, loam, sandy), weight residual are having kg/plot-1. In last, Harvest price is having Rs. /maund.

MULTICOLLINEARTIY ANALYSIS IN THE DATASET

Correlation matrix is calculated for checking the multicollinearity issue, but no multicollinearity is found in the dataset. Correlation matrix of the main factors for sunflower crops is calculated by using the Minitab software as in table 5.

Table 5: Correlation matrix for the variables

CM	VN	YED	PP	QTS	UR	DAP	ST	NW	UOM	WR	LC	ST	AKP	HP
VN	1													
YED	-0.13	1												
PP	0.02	0.00	1											
QTS	-0.15	0.18	-0.14	1										
UR	0.15	0.34	0.06	0.20	1									
DAP	-0.18	0.30	0.14	-0.21	0.36	1								
ST	0.39	-0.20	0.16	0.00	-0.16	-0.11	1							
NW	0.25	0.21	0.14	0.09	0.67	0.42	-0.10	1						
UOM	0.29	-0.22	0.65	-0.10	0.22	0.18	0.38	0.22	1					
WR	-0.20	-0.14	-0.24	-0.14	-0.21	0.01	-0.18	-0.32	-0.17	1				
LC	0.24	-0.34	0.17	-0.11	-0.22	-0.29	0.17	-0.30	0.32	-0.06	1			
ST	0.02	0.28	-0.05	-0.08	0.11	0.29	0.03	0.16	0.03	-0.04	-0.29	1		
AKP	-0.10	0.20	0.32	-0.30	0.15	0.43	-0.18	0.31	0.02	0.07	-0.15	0.25	1	
HP	0.17	-0.15	0.50	-0.27	0.30	0.11	0.03	0.37	0.67	-0.02	0.09	-0.09	0.07	1

Table 5, shows that no multicollinearity issue is found between predictors as all the values of the correlation coefficients are found to be less than 0.95 (Javaid et al., 2020). Thus the overall dataset is free from the issue of Multicollinearity.

BOXPLOT ANALYSIS

After diagnosis of multicollinearity, outliers are detected by using the boxplots in the dataset. The values outside the boxplot will be considered as outliers (Dawson, R. 2011). The boxplots are calculated for the predictors and response variables. Figure 2 - 7 represents the boxplots for the variables by using the Minitab software.

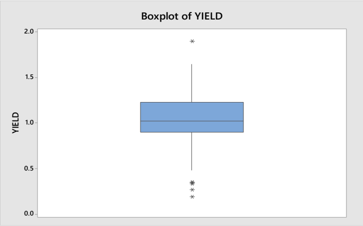

Figure 2: Boxplot of yield

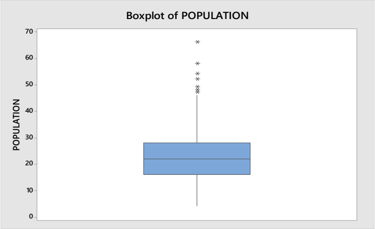

Figure 3: Boxplot of plant population

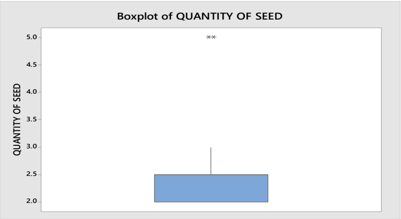

Figure 4: Boxplot QoS

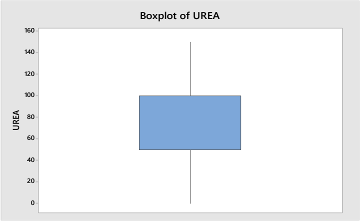

Figure 5: Boxplot of urea

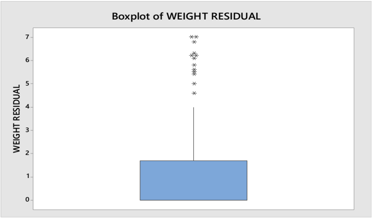

Figure 6: Boxplot of weight residual

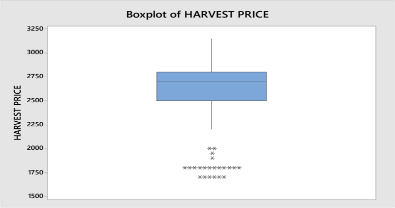

Figure 7: Boxplot of harvest price

From Figure 2 - 7, it is clear that the outliers are found in yield. The outliers are found in plant population according to the different villages. Outliers are also deducted in quantity of seed which shows that the quantity of seed is applied in variation for different villages. No outliers are found in urea. No outliers are deducted in last crop in the field. There is no outliers found in number of water well/ tube well. Outliers are observed in the weight residual. In the harvest price, the outliers are found in the dataset.

BEST MODEL SELECTION

After analysis of multicollinearity and outliers, the robust regression analysis is carried out for the best model selection due to presence of outliers.

Robust Regression Analysis

Three Method of M-estimators, Huber M- estimator, Hampel M- estimator and Tukey bisquare estimators are used to check the significance of factors. First Huber M- estimator is used by using the R software analysis. The results are carried out in term of Table 6 as follows.

Table 6: Result for the Method of Huber M- estimator

Factors	**Coefficient**	***p*-value**	**Significance**
Intercept	$4.026910e^{-01}$	$1.780207e^{-01}$	Non-significant
Village name	$3.475130e^{-03}$	$1.495577e^{-01}$	Non-significant
Plant Population	$9.974917e^{-03}$	$6.450133e^{-04}$	Significant
Qty of Seed	$1.258239e^{-01}$	$1.795653e^{-02}$	Significant
Urea	$3.155928e^{-03}$	$2.738321e^{-04}$	Significant
Dap	$3.824446e^{-03}$	$9.415086e^{-04}$	Significant
Soil type	$-1.281117e^{-02}$	$6.697652e^{-01}$	Non-significant
No. of Water well	$-2.440378e^{-02}$	$2.894936e^{-02}$	Significant

Usage of Machine	-2.440187e^{-01}	4.835849e^{-05}	Significant
Weight Residual	-1.229763e^{-02}	3.391567e^{-01}	Non-significant
Last Crop	-7.087166e^{-03}	4.433001e^{-01}	Non-significant
Seed treated	1.625088e^{-01}	4.608362e^{-03}	Significant
Attack pest	-2.780231e^{-02}	6.298947e^{-01}	Non-significant
Harvest price	6.319877e^{-05}	4.954744e^{-01}	Non-significant

From Table 6, the results of Huber M- estimator regression method. The final regression equation 1 with the significant factor can be observed as follows.

$\hat{Y}$ = 9.974917e^{-03}*Population*+1.258239e^{-01}*Qty of Seed* + 3.155928e^{-03}*Urea* + 3.824446e^{-03}Dap-2.440378e^{-02}*NoWater*-2.440187e^{-01}*StartMachine*+1.625088e^{-01}*Seedtreated* (1)

The regression output in equation (1) shows that the quantity of seed, urea, Dap, No water well / Tube well, population, start machine, seed treated predicators variable are statistically significant because the *p*-value are less than 0.05. The non-significant variables are village name, soil type, weight residual, last crop, Attackpest and harvest price because their *p*-value is greater than 0.05.

Another analysis is obtained by using the Hampel M-estimator in R studio Software. The result is given in Table 7 as follow

Table 7: Result for the Method of Hampel M- estimator

Factors	**Coefficient**	**p-value**	**Significance**
Intercept	4.422603e^{-01}	0.1445793478	Non-significant
Village name	3.252190e^{-03}	0.1829294995	Non-significant
Plant Population	9.020583e^{-03}	0.0022202242	Significant
Qty of Seed	1.222762e^{-01}	0.0231065725	Significant
Urea	2.986786e^{-03}	0.0006501072	Significant
Dap	3.437551e^{-03}	0.0032067598	Significant
Soil type	-1.342666e^{-02}	0.6591424859	Non-significant
No. of Water well	-2.360732e^{-02}	0.0368924080	Significant
Usage of Machine	-2.313348e^{-01}	0.0001352457	Significant
Weight Residual	-1.596647e^{-02}	0.2212026759	Non-significant
Last Crop	-7.807562e^{-03}	0.4046601368	Non-significant
Seed treated	1.629546e^{-01}	0.0050365402	Significant
Attack pest	-1.082517e^{-02}	0.8530406752	Non-significant
Harvest price	6.198943e^{-05}	0.5093371229	Non-significant

From Table 7, the results of Hampel M- estimator regression method can used in final model selection in equation (2).

$\hat{Y}$ = 9.020583e^{-03}*Population*+1.22262e^{-01}*Qty of Seed* + 2.986786e^{-03}*Urea* + 3.43751e^{-03}*Dap* -2.360732e^{-02}*No.Waterwell*-2.31338e^{-01}*StartMachine*+1.62954e^{-01}*Seedtreated* (2)

The regression output in equation (2) shows that the quantity of seed, urea, Dap, No water well / Tube well, population, start machine, weight residual, seed treated predicators variable are statistically significant because the *p*-value are less than 0.05. The non-significant variables are village name, soil type, last crop, Attackpest and harvest price because their *p*-value is greater than 0.05 (Lim et al., 2020).

The results of Tukey Bisquare M-estimator is carried out in R-studio software. The result are given in Table 8 as follows.

Table 8: Result for the Method of Tukey Bisquare M-estimator

Factors	Coefficient	p-value	Significance
Intercept	$4.075695e^{-01}$	$1.781250e^{-01}$	Non-significant
Village name	$3.471472e^{-03}$	$1.549810e^{-01}$	Non-significant
Plant Population	$9.964734e^{-03}$	$7.549013e^{-04}$	Significant
Qty of seed	$1.252138e^{-01}$	$1.995775e^{-02}$	Significant
Urea	$3.130367e^{-03}$	$3.584186e^{-04}$	Significant
Dap	$3.619801e^{-03}$	$1.932164e^{-03}$	Significant
Soil type	$-1.087220e^{-02}$	$7.206849e^{-01}$	Non-significant
No. of Water well	$-2.426401e^{-02}$	$3.188412e^{-02}$	Significant
Usage of Machine	$-2.434584e^{-01}$	$6.118377e^{-05}$	Significant
Weight Residual	$-1.167074e^{-02}$	$3.701305e^{-01}$	Non-significant
Last Crop	$-7.519528e^{-03}$	$4.217554e^{-01}$	Non-significant
Seed treated	$1.591743e^{-01}$	$6.071097e^{-03}$	Significant
Attackpest	$-2.562663e^{-02}$	$6.608228e^{-01}$	Non-significant
Harvest price	$9.359285e^{-05}$	$9.06747e^{-1}$	Non-Significant

From Table 8, Tukey Bisquare M- estimator regression equation can be obtained as follow in equation (3).

$\hat{Y}= 9.964734e^{-03}\text{Population} + 1.252138e^{-01}\text{QtyofSeed} + 3.130367e^{-03}\text{Urea} + 3.619801e^{-03}\text{Dap} - 2.426401e^{-02}\text{NoWater}-2.434584e^{-01}\text{StartMachine}+1.5917e^{-01}\text{Seedtreated}$ (3)

The regression output in equation (3) shows that the quantity of seed, urea, Dap, No water well / Tube well, population, start machine, weight residual, seed treated predicators variable are statistically significant because the *p*-value are less than 0.05 (Lodhi et al., 2023). The non-significant variables are village name, soil type, last crop, attackpest and harvest price because their *p*-value is greater than 0.05.

Ordinary Least Square Regression Analysis

Ordinary least square is used for comparison purpose among all the selected variables by using the Robust Regression analysis. Results are provided in the Table 9.

Table 9: Result for the method of Ordinary Least Square (OLS)

Factors	Coefficient	*p*-value	Significance
Intercept	$4.512e^{-01}$	0.131047	Non-significant
Village name	$3.134e^{-03}$	0.192671	Non-significant
Plant Population	$8.454e^{-03}$	0.003566 **	Significant
Qty of Seed	$1.198e^{-01}$	0.023947 *	Significant
Urea	$2.960e^{-03}$	0.000606 ***	Significant
Dap	$3.418e^{-03}$	0.002954 **	Significant
Soil type	$-1.394e^{-02}$	0.642030	Non-significant
No. of Water well	$-2.338e^{-02}$	0.035967 *	Significant
Usage of Machine	$-2.251e^{-01}$	0.000162 ***	Significant
Weight Residual	$-1.771e^{-02}$	0.168959	Non-significant
Last Crop	$-8.218e^{-03}$	0.373526	Non-significant
Seed treated	$1.604e^{-01}$	0.005074 **	Significant
Attackpest	$-1.952e^{-03}$	0.972963	Non-significant
Harvest price	$6.402e^{-05}$	0.489243	Non-significant

From Table 9, the final selected equation for OLS can be given as follows.

$$\hat{Y} = 8.454e^{-03}\mathit{Population}+1.198e^{-01}\mathit{Qty\ of\ Seed}+2.960e^{-03}\mathit{Urea} + 3.418e^{-03}\mathit{Dap}-2.338e^{-02}\mathrm{NoWater} - 2.251e^{-01}\mathrm{Start\ Machine}+ 1.604e^{-01}\mathit{Seedtreated} \quad (4)$$

The regression output in equation (4) shows that the quantity of seed, urea, Dap, No water well / Tube well, population, start machine, weight residual, seed treated predicators variable are statistically significant because the *p*-value are less than 0.05 (Javaid et al., 2020). The non-significant variables are village name, soil type, last crop, attackpest and harvest price because their *p*-value is greater than 0.05.

Ridge Regression

Ridge Regression is used on all predictors by taking the yield as response variable. The results are provided by using the R software analysis in Table 10 as follows.

Table 10: Result for the Ridge Regression analysis

Factors	Coefficient	p-value	Significance
Intercept	0.9874103150	0.0623	Non-Significant
Village name	-0.0006099031	0.0334	Significant
Plant Population	0.0009141979	0.0423	Significant
Qty of Seed	0.0385504600	0.0254	Significant
Urea	0.0010051588	0.0364	Significant
Dap	0.0010193120	0.0366	Significant
Soil type	-0.0188849260	0.0266	Significant
No. of Water well	0.0018494159	0.0366	Significant
Usage of Machine	-0.0367167181	0.0257	Significant
Weight Residual	-0.0090956431	0.0344	Significant
Last Crop	-0.0094672374	0.0379	Significant
Seed treated	0.0679623485	0.0279	Significant
Attackpest	0.0281391374	0.0223	Significant
Harvest price	-0.0000403473	0.0567	Non-Significant

From Table 10, Ridge Regression in equation (5) can be found by using all the significant factors as follows.

$$\hat{Y} =-0.0006\mathit{Villagename}+0.0009\mathit{Population}+0.03855\mathit{Qtyofseed}+0.0010\mathit{Urea}+0.00101\mathit{Dap}-0.0188\mathit{Soiltype}+0.0018\mathrm{NoWater}-0.0367\mathit{usageofmachine}-0.0090\mathit{WeightResidual}-0.0094\mathit{LastCrop}+0.0679\mathit{Seedtreated}+0.02813\mathit{Attackpest} \quad (5)$$

The regression output in equation (5) shows that the quantity of seed, urea, Dap, No water well / Tube well, population, start machine, weight residual, seed treated, village name, soil type, last crop, Attackpest and harvests price predicators variable are statistically significant because the *p*-value are less than 0.05 (Javaid et al., 2024). The non-significant variables is harvest price because its value greater than 0.05.

Comparisons of Robust Regression with OLS and Ridge Regression

Comparison of robust regression with OLS and Ridge Regression is done in Table 11 as follows.

Table 11: Comparison of Robust Regression with OLS and Ridge Regression

Significance variables Factors	**Coefficients**				
	Ridge	**OLS**	**Huber**	**Hampel**	**Bisquare**
Intercept	--------	------	------	------	------
Village name	-0.000609903	------	------	------	------

Population	0.000914197	8.454e^{-03}	9.9749e^{-03}	3.252190e^{-03}	9.964734e^{-03}
Qty of Seed	0.038550460	1.198e^{-01}	1.2582e^{-01}	9.020583e^{-03}	1.252138e^{-01}
Urea	0.001005158	2.960e^{-03}	3.1559e^{-03}	1.222762e^{-01}	3.130367e^{-03}
Dap	0.001019312	3.418e^{-03}	3.8244e^{-03}	2.986786e^{-03}	3.619801e^{-03}
Soil type	-0.018884926	-------	-------	-------	------
No. of Water well	0.001849415	-2.338e^{-02}	-2.4403e^{-02}	-1.34266e^{-02}	-2.426401e^{-02}
Usage of Machine	-0.036716718	-2.251e^{-01}	-2.4401e^{-01}	-2.36073e^{-02}	-2.434584e^{-01}
Weight Residual	-0.009095643	------	------	------	------
Last Crop	-0.009467237	------	------	------	------
Seed treated	0.067962348	1.604e^{-01}	1.6250e^{-01}	-7.80756e^{-03}	1.591743e^{-01}
Attackpest	0.028139137	------	------	------	------
Harvest price	-------	------	------	------	------

Non-significant factors from the final selected model are excluded on the basis of *p*-values that are greater than 0.05 (Javaid et al., 2019). Only two non-significant factor is excluded from the ridge regression analysis. While seven predictors are excluded by OLS in term of non-significant factors. From the three robust regression methods, seven significant factors are detected for the yield of sunflower while the seven non-significant factors are found from the three methods at 5% level of significance.

Efficient Model Selection

For the efficient model selection, Mean Square error (MSE) and Mean Absolute Percentage Error (MAPE), Akaike's information criteria (AIC) and Bayesian information criteria (BIC) are observed for each selected models through different estimators. The results are observed in Table 12 as follows.

Table 12: Result of MSE, MAPE, AIC and BIC for various regression estimators

	MSE	MAPE	AIC	BIC
OLS Method	0.0554	17.4044	-7.697213	36.53419
Ridge Regression	0.0543	22.5895	-7.789313	37.89419
Huber M-Estimators	0.0450	18.3647	-6.953132	37.27827
Hampel M-Estimators	0.0448	18.6205	-7.590992	36.64041
Bisquare M-Estimators	0.0450	18.4710	-6.981075	37.25032

From Table 12, On the basis of MSE, Hampel M estimator will be preferred because the MSE is minimum for the Hampel M-Estimator as compared to other techniques although the value of MSE and MAPE for OLS is minimum (Javaid et al., 2021) but it cannot be preferred as an efficient model because the OLS results are not effective in case of outliers (Gujrati; 2008). The lower value of Akaike's information criteria (AIC) and Bayesian information criteria (BIC) represents the better fitted model (Suhaeri et al., 2021). From the lower value of AIC and BIC, Hampel M estimator is preferred among all other estimators.

CONCLUSION

In this research, the efficient model is selected in different stages. Data analysis is started through multicollinearity analysis but no multicollinearity is found among predictors as all values are less than 0.95 (Javaid et al., 2020), then data is reanalyzed. Boxplot shows that there are presence of outliers in the dataset. Three M-estimators of robust regression (Huber M-estimator, Hampel M-estimator and Tukey bisquare M-estimator) are used in the second step to select the best model. Comparisons is made with OLS and Ridge regression of various robust estimates. In the third step, the results showed that Hampel estimators is efficient model to forecast based on MSE, MAPE, AIC and BIC values.

REFERENCES

1. Ahmad, H. H., Tahir, M. A., Sarwar, G., Sher, M., Aftab, M., Manzoor, M. Z., & Riaz, A. (2021). Growth and yield response of sunflower to organic amendments in aridisol. *Pakistan Journal of Agricultural Research*, *34*(1), 193.
2. Ahmad, M. I., Ali, A., He, L., Latif, A., Abbas, A., Ahmad, J., & Mahmood, M. T. (2018). Nitrogen effects on sunflower growth: a review. *Int J Biosci*, *12*, 91-101.
3. Ahmad, S., Ghaffar, A., Khan, M. A., & Mahmood, A. (2020). Evaluation of Different Production Systems in Combination with Foliar Sulphur Application for Sunflower (Helianthus annuus L.) under Arid Climatic Conditions of Pakistan. *Sarhad Journal of Agriculture*, *36*(4).
4. Andersen, R. (2008). *Modern methods for robust regression* (No. 152). Sage.
5. Bognár, P., Kern, A., Pásztor, S., Steinbach, P., & Lichtenberger, J. (2022). Testing the Robust Yield Estimation Method for Winter Wheat, Corn, Rapeseed, and Sunflower with Different Vegetation Indices and Meteorological Data. *Remote Sensing*, *14*(12), 2860.
6. Chatterjee, S., Hadi, A. S., & Price, B. (2006). Simple linear regression. *Regression Analysis by Example*,, 21-51.
7. Dawson, R. (2011). How significant is a boxplot outlier? Journal of Statistics Education, 19(2).
8. Dorugade, A. V. (2014). New ridge parameters for ridge regression. *Journal of the Association of Arab Universities for Basic and Applied Sciences*, *15*, 94-99.
9. Fernández-Luqueño, F., López-Valdez, F., Miranda-Arámbula, M., Rosas-Morales, M., Pariona, N., & Espinoza-Zapata, R. (2014). An introduction to the sunflower crop. *Sunflowers: Growth and Development, Environmental Influences and Pests/Diseases. Valladolid, Spain: Nova Science Publishers*, 1-18.
10. Gujarati, D. N. (2004). Basic econometrics (4th Ed.). New York, USA: The McGraw-Hill Companies
11. Gvozdenac, S., Milovac, Ž. Vidal, S., Crvenković, Z. L., Papuga, I. Š. Franeta, F., & Cvejić, S. (2022). Comparison of chemical and biological wireworm control options in Serbian sunflower fields and a proposition for a refined wireworm damage assessment. *Agronomy*, *12*(4), 758.
12. Javaid, A., Muthuvalu, M. S., Sulaiman, J., Ismail, M., & Ali, M. K. M. (2019, December). Forecast the moisture ratio removal during seaweed drying process using solar drier. In *AIP Conference Proceedings* (Vol. 2184, No. 1). AIP Publishing.
13. 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.
14. 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).Kadhim, H. M., & Abbas, S. H. (2020). Gene action for grain yield and some of its components in sunflower J. Plant Archives, 20(2), 7511-7518.
15. 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.
16. Javaid, A. (2024). Standard operating procedure for efficient model selection through hybrid of stepwise and robust regression analysis. *Remittances Review*, *9*(2), 3599-3615.

17. Khalifani, S., Darvishzadeh, R., Azad, N., & Rahmani, R. S. (2022). Prediction of sunflower grain yield under normal and salinity stress by RBF, MLP and, CNN models. *Industrial Crops and Products*, *189*, 115762.
18. 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).
19. 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.
20. Oilseed. 2019. Recent World Production, Markets, and Trade Reports, published by USDA, p. 70-115. Pakistan Bureau of Statistics. 2019. External Trade Division, Karachi. Radic, V., Vujakovic, M., Marjanovic-Jeromela, A
21. Perveen, S., Ahmad, S., Skalicky, M., Hussain, I., Habibur-Rahman, M., Ghaffar, A., & El Sabagh, A. (2021). Assessing the potential of polymer coated urea and sulphur fertilization on growth, physiology, yield, oil contents and nitrogen use efficiency of sunflower crop under arid environment. *Agronomy*, *11*(2), 269.
22. Pilorge, E. (2020). Sunflower in the global vegetable oil system: situation, specificities and perspectives. *OCL*, *27*, 34.
23. Qureshi, A. L., Gadehi, M. A., Mahessar, A. A., Memon, N. A., Soomro, A. G., & Memon, A. H. (2015). Effect of drip and furrow irrigation systems on sunflower yield and water use efficiency in dry area of Pakistan. *American-Eurasian Journal of Agricultural & Environmental Sciences*, *15*(10), 1947-1952.
24. Rousseeuw, P. J., & Leroy, A. M. (2005). *Robust regression and outlier detection*. John wiley & sons.
25. Saleem, M., Elahi, E., Gandahi, A. W., Bhatti, S. M., Ibrahim, H., & Ali, M. (2019). Effect of Sulphur Application on Growth, Oil Content and Yield of Sunflower. *Sarhad Journal of Agriculture*, *35*(4).
26. Shi, R., & Conrad, S. A. (2009). Correlation and regression analysis. *Ann Allergy Asthma Immunol*, *103*(4), S34-S41.
27. Suhaeri, M. E., Alimudin, A., Javaid, A., Ismail, M., & Ali, M. K. M. (2021, November). Evaluation of clustering approach with euclidean and Manhattan distance for outlier detection. In *AIP Conference Proceedings* (Vol. 2423, No. 1). AIP Publishing.
28. Tabassum, M. I., Aslam, M., Javed, M. I., Salim, J., Sarwar, M., & Rafiq, H. (2020). HYBRID DEVELOPMENT PROGRAMME OF SUNFLOWER IN PAKISTAN: A REVIEW. *Journal of Agricultural Research (03681157)*, *58*(3).
29. Vladimir, M. (2022). Introduction to the Special Issue Sunflower☆.