BEST STATISTICAL MODEL ESTIMATION FOR MAIZE YIELD IN KOTLI, AZAD KASHMIR

***Syed Masroor Anwar[1], Azhar Saleem[1], Muhammad Aslam[2], A.Q. Khan[3], M.N. Quershi[3], Syeda Maria Bukhari[4]**

[1]Department of Statistics,University of Azad Jammu and Kashmir,Pakistan
[2]Department of Statistics, Quaid-i-Azam, University, Islamabad, Pakistan
[3]Department of Mathematics,University of Azad Jammu and Kashmir,Pakistan
[4]Department of Health Sciences,University of Azad Jammu and Kashmir,Pakistan
*Corresponding author: masroor_anwar2001@yahoo.com

ABSTRACT: *This paper examines different factors that influence maize seed yield in Kotli, Azad Kashmir. Considering days of flowering, plant height, seed weight, branches per plant anddays to maturity as important factors (predictor variables)for maize seed yield (response variable)a regression analysis shows that three regressors; namely, the days of flowering, branches per plant and days to maturity have maximum contribution to increase maize seed yield in the area.*

1. INTRODUCTION

Maize being the highest yielding cereal crop in the world is of significant importance for country like Pakistan, where rapidpopulation growth has already out stripped the available food supplies. In Pakistan maize is third important cereal after wheat and rice. Maize accounts for 4.8% of the total cropped area and 3.5% of the value of agricultural output. Successful maize production depends on the correct application of production inputs that will sustain the environment as well as agricultural production.

Kiramat Khan *et al.*[1]conducted the study for the development of indigenous maize hybrids at the Cereal Crops Research Institute (CCRI) Pirsabak, Nowshera, Pakistan. The main objective was to evolve indigenous single-cross maize hybrids of high yield potential, white kernel and low to medium maturity for the environment of Khyber Pakhtoonkhua province. A large number of replicated field experiments, both on-station and on-farm, were carried out in 2005 and 2006 to evaluate the experimental hybrids. Grain yield, Stover yield, and maturity were among the important traits used in these investigations. A highest grain yield of 9.84 t ha^{-1} and a Stover yield of 30.56 t ha^{-1} with maturity earlier than other hybrids, including a leading maize hybrid of private sector (Pioneer-3025) were observed for one of the experimental hybrids, FRW-2 X FRW-8. This hybrid with a few exceptions was invariably found higher yielding and early maturing as compared to other hybrids included in the trials. The new experimental hybrid was officially named as "Kiramat". This study provided a sound basis for its approval by the 'Provincial Seed Council' and its registration by the 'Federal Seed Certification and Registration Department' for commercial cultivation.

Qureshi,A.H. *et al.* [2] conducted a study in order to explore farmers' practices in maize cultivation and consumption.The study was conducted covering major maize growing districts of Azad Jammu and Kashmir. Stratified random sampling technique was used to draw sample of 175 respondents and descriptive statistics was used to analyze the data. He foundthat majority (59 percent) of the farmers were growing traditional verities. On an average the farmers were using 3.41 kg per kanal seed. It was also observed that use of chemical fertilizer was very low in quantity terms and found no trend ofapplying pesticides. Average yield of 401 kg per acre was observed. About eighty two percent of the maize production isutilized by the former himself; six percent reserved for seed and just ten percent of the production is sold to fellow farmers or in the market. Maize is the main crop grown during kharif season in Azad Jammu and Kashmir. Overall more than 80 percent of the cultivated land is allocated to maize crop during kharif. However, some inter-district variation was observed in the sample area. For example in districtBagh maximum (88 percent) area is allocatedto this important crop.

Oyewo*et al.* [4] conducted study using a cross sectional data obtained through a multistage sampling technique.This study estimated the technical efficiency of maize producing-farmers in Oyo State, Nigeria and further examined the factors that determine thedifferential in efficiency index. A multistage sampling technique was used to select 120 maize farmers in the study area. Data were collected and subjected to inferential statistics; stochastic frontier production model was used in the analysis to determine the relationship between the maize output and the level of input used in the study area. The empirical results revealed that farm size and Seed were statistically significant at 10% and 1% level respectively in the study area. The estimated gamma parameter in the study area indicates that 12% of the total variation in maize output is due to the technical inefficiencies.

Sadiq *et al* [5] conducted study in selected areas of agro ecological zones-2 and 3 (district Muzaffarabad, Poonch and Samahni - Bhimber) of Azad Jammu and Kashmir in order to estimate the technical efficiency through stochastic Cobb-Douglas production frontier. The mean technical efficiency index was found as 68 % with minimum of 30 % and maximum of 94 %. They concluded that farmers with more age and more education were technically less efficient while farmers with larger farm size and having greater contact with extension agents were found more efficient. Tractor hours, use of farm yard manure and labor-hours contribute significantly to maize productivity while use of any chemical fertilizer was found insignificant in maize productivity. Farm size and close contact with the extension agents have shown negative impact on inefficiency effect, indicating that farmers with large farm size with more contact with extension agents were technically more efficient than their counterparts [5].

The most important issue for the analysis of this kind of data is the selection of suitable model; otherwise results may give erroneous impression about the information provided. To see the feasibility of the proper statistical technique we check the assumptions i.e. Normality, linearity, hectosecdasticity, autocorrelation and Multicollinearity. For this purpose residual plot, variance inflation factor (VIF), Durban Watson test etc. is employed for the particular regression model that is suitable for our data. Since an outlying observation may be merely an extreme manifestation of the random variability inherent in the data and there is a chance of an error in calculating or recording a numerical result.So, outliers are detected by applying different techniques.

2. MATERIAL AND METHODS

Maize data was obtained from National Agricultural Research Center, Islamabad (NARC). The experimental material consists of maize genotypes arranged in Randomized Complete Block Design. In this study maize data related to morphological plant characteristics such as Days of flowering(X_1), Plant Height (X_2), Seed Weight (X_3), Branches per plant (X_4), Days to Maturity (X_5), are taken as predictor variables and Seed Yield (Y)as the response variable. Different regression assumptions related to residuals; Normality of residuals, Linearity of regression model, Homoscedasticity or equal variance of errors, Problem of autocorrelation and Multicollinearity are described also different methods of detection of outliers are described.

To study the outlying fixed traits (X observations), method of Leverage values isemployed whereas for outlying Y observation Studentized deleted residuals and for influential observations Cook's Distance and DFFITS are used.

For selecting a best model various criteria such as best subset regression and Stepwise Regression are used.Using the subset regression procedure and other criterions such as R^2, R^2_{adj} and Mallows C_p statistic we got information about the best model. On the basis of two criterions, namely the subset regression and the stepwise regressiongave the best model having three regressors.

2.1 Autocorrelation: Durban Watson test is used to detect the autocorrelation in the data. The residuals are estimated for errors in the model are assumed to be independent. The Durban Watson test checks for a sequential dependence of error terms. Mathematical formulation of the test is:

$$d = \frac{\sum_{u=2}^{n}(e_u - e_{u-1})^2}{\sum_{u=2}^{n}(e_u)^2},$$

where$d$ is the residual and e_u denotes the errors.

2.2 Multicollinearity:Variance Inflation Factor (VIF) is applied to check the existence of Multicollinearity in the Maize data. It is calculated using the formula $VIF = \frac{1}{1-R_j^2}$,whereR_j^2 is the squared multiple correlation coefficient. A value $VIF > 10$ indicates the presence of multicollinearity (Drapper and Smith, 2003) [8].

2.3 AdjustedR^2: It is denoted byR^2_{adj} and defined by the equation:

$$R^2_{adj} = 1 - \left(\frac{n-1}{n-p}\right)(1 - R^2_p)$$

Where p is the number of regressors included in the model [8].

2.4 R^2_p Criterion: It is denoted by$R^2_p$and defined as

$$R^2_p = \frac{SSR_p}{SST_0},$$

whereSSR_p is the regression sum of squares of$p$ regressors included in the model (Drapper and Smith, 2003) [8].

2.5 Mallows C_p Criterion: Mallows (1973,1995) applied $C_p$statistic for the indication of the best subset regression model and is defined as:

$$C_p = \frac{RSS_p}{s^2 - (n-2p)},$$

Where RSS_p is the residual sum of squares from the model containing p parameters in the model [9].

2.6 Cook's Statistic: Cook [7] proposed that the influence of the i^{th} data point be measured by squared scaled distance

$$D_i = \frac{(\hat{Y} - \hat{Y}_{(i)})'(\hat{Y} - \hat{Y}_{(i)})}{ps^2},$$

Where$\hat{Y} = Xb$ is the usual vector of the predicted values, while$\hat{Y}_{(i)} = Xb_{(i)}$ is the vector of predicted values from a least squares fit when the i^{th}data point is deleted, where $b_{(i)}$ is the corresponding least squares estimator.

2.7 The DFFITS Statistics:These statistics are cousins of Cook's influence measures. They are defined by Belsley, kuh and Welsch (1980) as:

$$DFFITS = \left[\frac{(b - b_{(i)})'X'X(b - b_{(i)})}{s^2_{(i)}}\right]^{\frac{1}{2}}, = \left[\frac{D_i p s^2}{s^2_{(i)}}\right]^{\frac{1}{2}},$$

Where $D_i$is the Cook's statistic and $S^2_{(i)}$ is the estimate ofσ^2 obtained after deletion of the contribution of the i^{th} residual.

3 RESULTS AND DISCUSSIONS

3.1 Normality of Residuals: Figure-1 present normal probability plot of residual.All the points are reasonably close to the straight line which reveals that the residual follows normal distribution [8].

3.2 Linearity and Homoscedasticity of error variance:Studentized residuals verses fitted values of $\hat{Y}_{(i)}$in Figure-2 shows that there is a random scatter of points along the horizontal line which means stability of the linear model. The randomness of points in figure-2 also shows that there is homoscedasticity of the error variance [8].

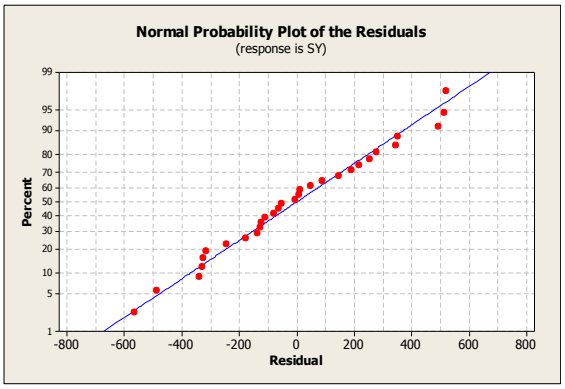

Figure-1:Normal Probability Plot of residuals

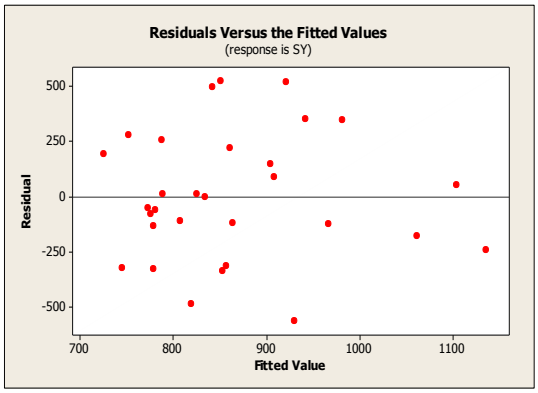

Figure-2:Residuals versus the fitted values

3.3 No Autocorrelation between the error: As the calculated value of Durban Watson Test is 2.07 which is close to 2 showing no presence of autocorrelation in the data [8].

3.4 Problem of Multicollinearity:For the detection of multicollinearity we used variance inflation factor(VIF) criterion. From table-1 below we see that all the calculated values of (VIF) are less than 10 indicating absence of multicollinearity in the data [8].

Table-1 Variance Inflation Factor for Maize data

Predictor	P-VALUE	Co-efficient	VIF
Constant	0.531	663	0.0
X_1	0.633	2.909	1.4
X_2	0.835	0.659	1.4
X_3	0.175	176.3	2.4
X_4	0.650	-27.88	1.1
X_5	0.605	-3.131	1.1

Days of flowering $= X_1$, Plant height $= X_2$, Seed Weight $= X_3$, Branches per plant $= X_4$, Days to Maturity$= X_5$

3.5 Detection of Outliers: To identify the outliers in the data different criterion such as Leverage value, Mahalanobis Distance, Studentized residuals, DIFFITS and Cook's Distance were applied their calculated values aregiven in Table-4.

3.6 Leverage Value: Leverage values of hat matrix$h_{06} = 0.412219$ of observation 06 exceeds the criterion of twice the mean leverage value,$\frac{2p}{n} = 2(0.1999) = 4$. So we conclude that the previous mentioned observation has outlying fixed traits [8].

3.7 Studentized deleted residuals: To identify outlying Maize yield, we apply Studentized deleted residual by considering the tail area 5% on both sides of the $t-$distribution and compare these large absolute studentized deleted residuals with$t_{(0.95;23)} = 1.7170$. Observations are shown in Table-4. Note that the genotypes (observations) "Maize (observation # 23)", "Maize (observation # 24)"and "Maize (observation # 26) and Maize (observation # 27)" has the largest absolute Studentized deleted residuals. Now to check whether these observations are influential observations, Cook's Distance and DIFFITS were observed. Since all observations are less than the F-value; $F_{(0.95;7;23)} = 3.40$. It indicates that there influence is not enough strong to proceed for remedial measure [8].

3.8 Model selection: For the selection of suitable model Best Subset regression and Stepwise regression were applied.In Best Subsets Regression the magnitude of different statistics i.e.,$R^2{}_p$, $R^2{}_{adj}$, MSE and Mallow'sC_{p_i} helped us in finding best fitted model.From the figure-3 and table-3 $R^2{}_p$ increases when we include the three predictors i.e. Days of flowering (X_1), Branches per plant (X_4) and Days to Maturity (X_5) in the model. Then $R^2{}_p = 51.1$ indicate that 51.1% variability in the Maize yield production is because of the Days of flowering, Branches per plant and the Days to Maturity. The model is good with these explanatory variables according to the $R^2{}_p$ criterion [8].

Similarly for the best $k = 3$ the calculated values of $R^2{}_{adj}$ are shown in table-3. It is observed from table-3 and figure-5 that the three predictors such as Days of flowering (X_1), Branches per plant (X_4) and Days to Maturity (X_5) with$R^2{}_{adj} = 42.7\%$ are considerable factors which significantly affect the Maize seed yield. Now from Mallow's C_p criterion we see from table-3 and figure-4 that the value of $C_p$is so small and close to number of parameters$p$when three predictors i.e. Days of flowering (X_1), Branches per plant (X_4) and Days to Maturity (X_5) are included in the model with $C_p = 3.1$that is close to p. Hence according to Draper and Smith (2001) Mallow'sC_p criterion select the best model with(X_1),(X_4) and(X_5) as explanatory variables.

From the table-2 we find that Stepwise regression select the model with the variables i.e. Days of flowering (X_1), Branches per plant (X_4) and Days to Maturity (X_5) that can significantly affect the Maize seed yield in Kotli, Azad Kashmir with $R^2 = 51.5\%$.

CONCLUSIONS AND RECOMMENDATIONS

The above results show that both the methods; Best Subset Regression and Stepwise Regression the variables imply that the Days of flowering (X_1), Branches per plant (X_4) and Days to Maturity (X_5)has significant effect on the Maize production in district Kotli, Azad Jammu &Kashmir.It is recommended that similar studies be conducted in other

districts of Azad Kashmir and in all other provinces of Pakistan to find variables that impact Maize production .

Table 2: Model fitting using Best Subset and Stepwise regression

Best subset regression	$\hat{Y} = \beta_0 + \beta_1 X_1 + \beta_4 X_4 + \beta_5 X_5$			
	β_0	β_1	β_2	β_3
Coefficients(p-value)	845(0.00)	125(0.00)	7.23(0.039)	8.29(0.00)
$R_p^2 = 51.1\%, R_{adj}^2 = 42.7\%$, $MSE = 144$, p −value=*0.00, Mallow's* C_p =3.1				
Stepwise Regression	$\hat{Y} = \beta_0 + \beta_1 X_1 + \beta_4 X_4 + \beta_5 X_5$			
	β_0	β_1	β_2	β_3
Coefficients(p −value)	845(0.00)	125(0.00)	7.23(0.039)	8.29(0.00)
$R_p^2 = 51.1\%, R_{adj}^2 = 42.7\%$, $MSE = 144$, p −value=*0.00, Mallow's* C_p =3.1				

Table 3: Results of subset regression criteria.

p_{-1}	R^2	R_{adj}^2	C_p	S	X_1	X_2	X_3	X_4	X_5
1	41.9	39.5	6.4	152.46				X	
1	34.5	31.8	10.0	160.52					X
2	26.5	23.5	13.8	168.83		X		X	
2	13.4	8.6	17.2	176.00				X	X
3	46.4	38.4	1.6	137.79	X	X		X	
3	51.1	42.7	3.1	144.40	X			X	X
4	38.4	33.9	5.3	147.61	X	X		X	X
4	34.4	29.6	7.2	152.32	X	X	X	X	
5	56.6	43.4	2.1	136.30	X	X	X	X	X

Days of flowering= X_1, Plant height= X_2, Seed Weight= X_3, Branches per plant= X_4, Days to Maturity= X_5.

Table-4 Diagonal elements of Hat matrix h_{ii}, DFFITS, Studentized delete residuals d_i and cook's distance

Obs. #	d_i	h_{ii}	Cook's Dis.	DFFIT
1	0.49963	0.162049	0.008306	0.21972
2	1.04424	0.293013	0.075039	0.67226
3	0.89194	0.211146	0.035795	0.46145
4	-0.16881	0.104999	0.000581	-0.05782
5	0.03702	0.137188	0.000038	0.01476
6	-0.31843	0.412219	0.012313	-0.26667
7	-0.63328	0.238913	0.021519	-0.35481
8	0.73426	0.338265	0.046832	0.52497
9	0.19006	0.363744	0.003586	0.14371
10	-0.36821	0.148856	0.004099	-0.15398
11	-0.45664	0.151987	0.006441	-0.19332
12	0.76409	0.202618	0.025162	0.38517
13	-0.40296	0.12027	0.003834	-0.14899
14	-0.22159	0.225071	0.002475	-0.11942
15	0.03316	0.135812	0.00003	0.01315
16	-0.01577	0.087702	0.000004	-0.00489
17	-0.45768	0.262496	0.012849	-0.27305
18	-0.9176	0.311063	0.063781	-0.61658
19	-1.10988	0.193364	0.048744	-0.54341

20	-1.13289	0.099827	0.023445	-0.37727
21	-1.18433	0.227562	0.067734	-0.64282
22	-1.14831	0.192961	0.051858	-0.56149
23	-1.96174	0.090057	0.056745	-0.61715
24	-1.85203	0.249883	0.172929	-1.06894
25	0.30163	0.17286	0.003294	0.13789
26	1.84138	0.220214	0.145133	0.97854
27	1.93903	0.222065	0.160429	1.03599
28	1.773	0.073673	0.038252	0.50001
29	1.24106	0.190603	0.05912	0.60225
30	1.1938	0.159518	0.044296	0.52008

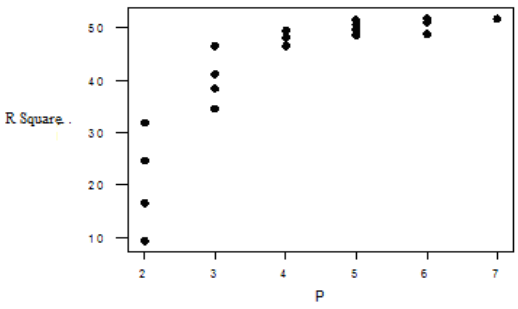

Figure-3 plot of R Square values versus P

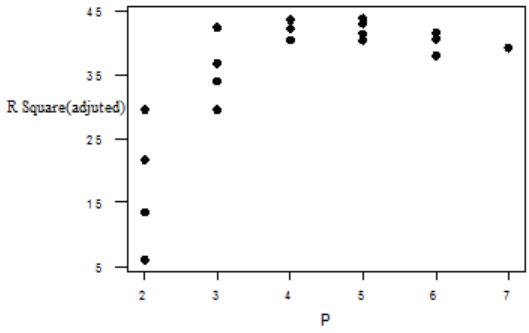

Figure-5 plot of Adjusted R Square values versus P

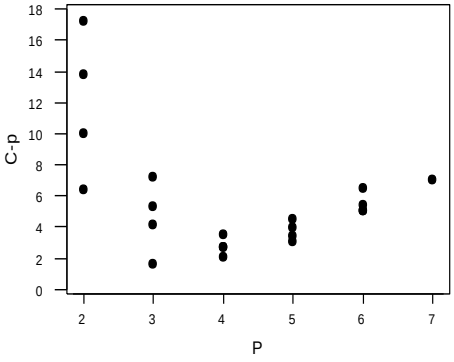

Figure-4 plot of Cp values versus P

REFERENCES:

[1].Kudi, T. M., Bolaji, M., Akinola M. O. and Nasa'I D. H."Analysis of adoption of improved maize varieties among farmers in Kwara State, Nigeria", *International Journal of Peace and Development Studies*,**1**(3),8-12(2011).

[2].Qureshi, A. H., Abbasi,S.S, Sharif,M. and Malik,W. "Farmer Practies of Maize production and consumption in Azad Jammu and Kashmir", *Pak. J. Agri. Sei.*,**39**(4),287-291(2002).

[3].Khan, K., Sher, H., Iqbal, M. and Al-Qurainy,F."Development and release of indigenous maize hybrids to enhance maize yield in khyber-Pakhtoonkhua province of Pakistan",*African Journal of Agricultural Research*, **6**(16),3789-3792(2011).

[4].Oyewo, Isaac. O."Technical efficiency of maize production in Oyo state",*Journal of Economics and International Finance*,**3**(4), 211-216(2011).

[5].Sadiq, G., Haq, Z. U., Ali,F., Mahmood,K., Shah,M. and Khan,I. "Technical Efficiency of maize farmers in various ecological zones of AJK",*Sarhad J. Agric.*,**25**(4),607-610(2009).

[6].Zafar,M., AbbasiM,K., Khaliq,A. and Rehman,Z.U. "Effect of combining organic materials with inorganic phosphorus sources on growth, yield, energy content and phosphorus uptake in maize at Rawalakot Azad

Jammu and Kashmir", Pakistan, *Archives of Applied Science Research*,**3**(2),199-212(2009).

[7].Cook,R.D."Detection of influential observations in linear regression", *Technometrics*,**19**,15-18(1977).

[8].Draper,N.R. and H. Smith, "Applied Regression Analysis", *Wiley Series in Probability and Statistics;* 3rd edition(2003).

[9].Mallows, C. L."*Some comments on* C_p*",Technometrics*, **15**,661-675(1973).

[10].Mallows, C. L., "More comments on C_p",*Technometrics*, **37**, 362-37(1973).