Global Jour. of Maths. & Stat.
Volume 2, Number 2 (2010) 103-112

STATISTICAL ASSESSMENT OF MODELS FOR DAURO CEREAL CROP

ABDULLAHI. F. BUSARI and **USMAN ABUBAKAR**
Department of Mathematics and Statistics
Federal University of Technology, Minna, Niger State, Nigeria

Abstract : Statistical tools play greater roles in the assessment of models for all items or variables for the purpose of prediction or drawing inferences about the goods or variables. Leaves, Plant height, Number of tillers, Root, Leaf s area etc are some of the plant attributes that are of great concern to Agricultural researchers because of their immense but varying role they play in plant growth which subsequently affect the plants. In this paper; we wish to establish the best statistical model for Dauro cereal crop considering the plant height, leaf's areas, number of leaves and number of tiller's contributions to the model using Ameriya prediction criterium. It was observed that dropping the number of leaves would lead to obtaining the best of the models for Dauro Crop. It was also discovered that the worst model would be obtained when leafs square area was excluded. Thus, the best of the models for Dauro Crop is given by the equation $Y = -0.09519 + 0.03125\ X_1 + 0.000264\ X_2 + 0.004795\ X_4$
It was also discovered that the worst model would be obtained when leafs square area as plant attributes was dropped. It is therefore significant that when number of tiller is excluded the best model for Dauro crop will be obtained.

Introduction

It has been convincingly proved that some arable crops solely depend on plants attributes such as plant height, number of Tillers, number of Laves, stalk numbers, spacing, leafs area and varieties. What one crop relies on or counts as significant attribute that contribute in no small measures to the yield is what reduces the yield in another crop. For example, it was proved that in Maiwa crop plant height is insignificant plant attribute to get the best model but that number of leaves is one of the significant plant attributes that contributes in no small measure, to its best model. (Busari. A.F; Abubakar .U; and Cole A.T; 2010) that is inclusion of Plant height leads to poor model. According to Rizzi Laura (2008) stated that the followings lead to error in the matrix. if (i) Relevant variables are excluded (ii) Irrelevant variables are excluded (iii) An incorrect functional form is used That is some models may not require linear equations to fit the data set but some other forms. Similarly, Ogunremi (1970) Exclaimed that Pod number/unit area is an important covariate that determines the yield in Pod producing crops. In fibers crops such as KENAF, it is the plant height that determines the significant yield (Baker, 1970). While in cotton, it is the number of bolls that determines the yield (Garder and Tucker, 1967). However, there is no point

Keywords : Dauro crop, Tillers, Plant height, Prediction criterium, and Regressor.

doubting the fact that the leaves stem and roots are significant plant organs/attributes of a plant. Hence, we should accept a strong relationship between these attributes and the yield. The relationship within these attributes is another source of variation that could affect having a good fit. If there is correlation between two or more covariates, then there is possibility of MULTICOLLINEARITY. And this affects the goodness-of-fit of a model. These and other necessary factors shall be considered to establish the best model for Dauro crop.

Method of data collection

The data used in this paper was secondary data obtained from the institute for Agricultural Research (IAR, Ahmadu Bello University, Samaru, Zaria. According to the Institute documentary, three varieties (V_1, V_2 and V_3) were planted in different treatment combinations using four different spacing (S_1, S_2, S_3, and S_4). In the experiment a plot constitute six (6) ridges each of which is of size 6m × 6m and a replication of plots of size 69m × 4m. From the two central ridges, data were collected while the first two ridges and the last two ridges serve as discards. Three seeds per hole were planted on top of the ridges and well covered with top soil to prevent rodents from eating the seeds. Fertilizer N and P_2O_5 were applied in blanket form. Immediately after the germination, thinning was done and vet ox 85 was spayed to prevent insects and hoeing was frequently and thinning done. Each net plot was numbered according to the number of tillers. The data is as shown in appendix-1.

Literature Review

Regression Analysis

This is a technique that is used establish if there exist linear relationship between the dependent variable o regressors, Predictors or covariates while, the independent variable could be referred to as a response to as a response or yield. Regression analysis is generally classified into two that is Simple and multiple linear regression analyses, having the same assumptions about the error term. In this paper multiple regression analysis would be used because we are considering more than one independent variable this will lead us to obtaining a probabilistic model.

According to Sarah P. Otto and Tony day (2007) Stress that modeling in Ecology and Evolution, the probabilistic model contains a random component which accounts for the error of the deterministic component. This random component accounts for measurable and immeasurable variations of the model. Hence the regression analysis model is probabilistic in nature because it includes the error term. The model is as written below.

$$y_{ij} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_k x_k + e_{ij} \quad (1)$$

Where y_{ij} = is the response or yield or dependent variable

x_i = is the predictor, regressor, covariate or independent variable

β_0 = is the intercept on the y — axis

β_i = is the i^{th} coefficient of regression

e_{ij} = is the error term

The β_0 and β_i are the unknown population parameters which can be estimated using the least square method as stated below or matrix method

From the equation above, the intercept on the y-axis is –0.09209 while the yield increases by

Least square method for Simple Linear Regression

$y_{ij} = \beta_0 + \beta_1 x_1 + e_{ij}$ (2)

Where $e_{ij} = y_i - \beta_0 - \beta_1 x_1$ and $e_{ij}^2 = (y_i - \beta_0 - \beta_1 x_1)^2$

$$\Rightarrow \sum_{i=1}^{n} e_{ij}^2 = \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_1)^2 = SSE$$

Minimizing the sum of squares of error and differentiating with respect to fa yielded the following equations

$$\frac{\partial}{\partial \beta_i} \sum_{i=1}^{n} e_{ij}^2 = \frac{\partial}{\partial \beta_i} \sum_{i=1}^{n} (y_i - \beta_0 - \beta_i x_i)^2 = 0$$

Solving for $\hat{\beta}_0$ and $\hat{\beta}_1$ derived above simultaneously resulted to the following estimates for

$\hat{\beta}_0$ and $\hat{\beta}_1$ $\hat{\beta}_0 = \hat{y} - \hat{\beta}_1 \bar{x}$ and $\hat{\beta}_1 = \dfrac{\sum_{i=1}^{n} x_i y_{i-n\bar{x}\bar{y}}}{\sum_{i=1}^{n} x_i^{2-n\bar{x}^2}}$

Least square method for Multiple Regression Analysis

For Multiple regression analysis, the K independent variables are assumed to be linearly related to the independent variable or response. The appropriate model is

$y_i = \beta_0 + \beta_i x_i + \beta_2 x_2 + \beta_3 x_3 + ... + \beta_k x_k + e_{ij}$ (3)

The least square method could also be used to estimate the population parameters by minimizing the sum of squares of error and differentiating with respect to each population parameter

$e_{ij} = y_i - \beta_0 - \beta_1 x_1 - \beta_2 x_2 - \beta_3 x_3 - ... - \beta_k x_k$ (4)

$e_{ij}^2 = (y_i - \beta_0 - \beta_1 x_1 - \beta_2 x_2 - \beta_3 x_3 - ... - \beta_k x_k)^2$ (5)

$$\frac{\partial}{\partial \beta_i} \Sigma e_{ij}^2 = \frac{\partial}{\partial x} \Sigma (y_i - \beta_0 - \beta_1 x_1 - \beta_2 x_2 - \beta_3 x_3 - ... - \beta_k x_k)^2 \quad (6)$$

$$\frac{\partial}{\partial \beta_i} \Sigma e_{ij}^2 = -2\Sigma (y_i - \beta_0 - \beta_1 x_1 - \beta_2 x_2 - \beta_3 x_3 - ... - \beta_k x_k)^2 = 0$$

$$\frac{\partial}{\partial \beta_i} \Sigma e_{ij}^2 = -2\Sigma x_i (y_i - \beta_0 - \beta_1 x_1 - \beta_2 x_2 - \beta_3 x_3 - ... - \beta_k x_k)^2 = 0 \quad (7)$$

Differentiating equation (7) result into (K+l) equations with (K+l) unknown parameters which can be solved simultaneously for the values of (K+l) unknown parameters i.e. β_0, β_1, β_2, β_3, ..., β_k.

Similarly, Matrix notation could be used to obtain the estimate of the population parameters or the coefficients of the regression. This is done as follows.

In matrix form equation (2) can be written as

$\underline{Y} = \underline{X}\underline{\beta} + \underline{\varepsilon}$ where

$$\underline{Y} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_k \end{pmatrix} \quad \underline{X} = \begin{pmatrix} x_{11} & x_{12} & x_{13} & \cdots & x_{1k} \\ x_{21} & x_{22} & x_{23} & \cdots & x_{2k} \\ x_{31} & x_{32} & x_{33} & \cdots & x_{3k} \\ & & \vdots & & \\ x_{n1} & x_{n2} & x_{n3} & \cdots & x_{nk} \end{pmatrix}; \varepsilon \sim \text{iid } N(0, \delta_i^2)$$

$$E(\varepsilon) = 0 \text{ and } \operatorname{cov}(\varepsilon) = \begin{pmatrix} 1 & 0 & 0 & \cdots & 0 \\ 0 & 1 & 0 & \cdots & 0 \\ 0 & 0 & 1 & \cdots & 0 \\ & & \vdots & & \\ 0 & 0 & 0 & \cdots & 1 \end{pmatrix}$$

$\underline{X}$ is of full rank of K+l; estimate of $\underline{\beta}$ is unbiased of β while the estimate of δ^2 is as reviewed below

$$E\left(\delta^2\right) = \frac{\delta^2(n-r)}{r} \cdot \frac{r}{(n-r)} = \delta^2 \text{ which is the unbiased estimate of } \delta^2$$

According to Rizzi Laura (2008), the assumptions that guide the linear model are related to the distribution of (i) error matrix (ii) independent variables (iii) unknown population matrix. If any of the assumption is wrong then there would be problems with the assumptions that relates to (a) distribution of errors (b) choice of regressors or independent variables (c) the estimates of the population parameters. That is (i) if the e_{ij} is not identically independently distributed as normal variate then the inference produced shall only be valid asymptotically, (ii) if the variance of e_{ij} terms are not constants or the same (homoscedasticity), then the error terms are heteroscedasticity, which occurs in cross-section data (iii) if the error terms are pairwise corrected, i.e. $E\{\varepsilon_i, \varepsilon_j\} \neq 0$; $i \neq j$. This happens in time series data, it cause autocorrelation. Furthermore Rizzi Laura enumerates actions that usually lead to a bad model, if (a) irrelevant independent variables are included (b) relevant independent variables are excluded (c) incorrect functional form is used to fit the data (d) matrix of $\underline{X}$ has less than full column rank (e) the independent variables are correlated or correlated with the error term.

This last point causes multicollinearity. The effect of which is always high coefficient of determination R^2 and the estimates of coefficient of regression are always insignificant. Multicollinearity also leads to high correlation coefficient between the independent variables and high variance inflation factors (VIF).

The existence of multicollinearity can be remedied by removing one or more independent variables observed to be causing multicollinearity, increase the sample size or transform the equation model. He further suggested a prediction criterion called AMEMYA PREDICTION CRITERIUM (PC).

This is used to evaluate the goodness of fit of a model i.e.

Prediction criterion $= \frac{RSS\{n+(K-1)\}}{\{n-(K-1)\}}$ where

RSS is the regression sum of square; n is the observation's number; (k-1) is the number of independent variables in the model.

Assessing the Model

The followings are in addition to those mentioned above, methods of assessing a model (i) standard error of the estimate (ii) t- test of the slope i.e. β_i; (iii) coefficient of determination; R^2 (iii) F- Ration or P- value in ANOVA

These are based on the sum of squares of error (SSE). As we have already assumed that e_{ij} is independently and identically distributed as a normal variate with mean zero and variance δ^2 i.e. $\varepsilon_i \sim$ iid N $(0, \delta^2_\varepsilon)$

If δ_ε—large then some errors will also be large which implies that the model is poorly fitted. But if δ_ε — is small this implies that some of the errors will be closed to zero; this implies a good fit model.

The standard error is estimated as If $\delta_\varepsilon = \sqrt{\frac{SSE}{n-2}}$, where SSE is the sum of square of error and n is the number of observation. If the error is normally distributed the test statistics is student-t-distributed with n-2 degrees of freedom. This is used to prove or established the existence of a linear relationship between (the independent and dependent variables) two or more covariates.

Table-I : Analysis of Variance for Regression Analysis

Source of Variation	Degrees of freedom	Sum of square	Mean sum of square	F-ration	P-Value
Regression	K	SSR	SSR/K	MSR/MSE	
Residual error	n-k-1	SSE	SSE/n-k-1		
Total	n-1	SST			

A large value of F-ration indicates that a significant proportion of the variation in Y is explained by the regression and that the model is valid. Similarly a small value of F-ration implies that most of the variation in Y is explained. i.e. if F-ration calculated $> F_{(n_1,n_2,\infty)}$ reject H_0 otherwise accept H_0; if P-value $\leq \infty$ (level of significance) or the critical region, reject H_0. It evident that the F-test is performed when there is more one independent variable.

Data Presentation and Analysis

The data in Appendix 2 was used to obtain the models for the assessment. A multiple regression analysis using Minitab 14 statistical software as in Appendix-1 gives the models, the model for Dauro crop considering all the four independent variables was considered here without dropping any of the variables

$$Y = -0.09209 + 0.03008\ X_1 + 0.0003453\ X - 0.001077\ X_3 + 0.005237\ X_4 \quad (8)$$

Sum-of-squares (RSS) = 0.003670; SD of residuals = 0.01564; R squared = 0.2233;

F = 1.0784; The P value is 0.4018; VIF < 5 in all cases, n = 20, K – 1 = 4;

$F_{(n_1,n_2,\infty)} = F_{4,15,0\cdot05)} = 3.06$

From the equation above, the intercept on the y-axis is –0.09209 while the yield increases by 0.03008 for every unit increase in plant height and by 0.0003453 for every unit increase in the number of leaves. Similarly, about 0.005237 increases in the yield was observed for every one unit increase in leafs square area and 0.001077 decreased in yield was noticed for every unit increase in the number of tillers. The R^2; coefficient of determination obtained showed that about 22% of the total variation due to regression, while the remaining percentage was unexplained. This is not good enough if a higher percentage could be due to residual error. The F- value and the P-value confirmed the invalidity of the model because F- calculated is less than the value of F-table =3.06 and the P-value is also greater than 0.05, the level of significance. Hence the model has goodness of fit. Although the VIF is less than five in all the cases and the R^2 is also less than 0.75. These two values confirmed the non-existence of multicollinearity among the variables (independent); the t-test values also proved that there is no sufficient difference to reject the null hypothesis of no linear relationship. This is the report when all the independent variables were considered. The Ameriya Prediction criterion is calculated as follows.

$PC_i = \frac{RSS\{n+(k-1)\}}{\{n-(k-1)\}}$ Where RSS is the Regression sum of square, n is the number of observations, (K – l) is the number of independent variables in the model and i =1, 2, 3, 4.

$$PC_1 = \frac{RSS\{n+(k-1)\}}{\{n-(k-1)\}} = \frac{0.003670x24}{16} = 0.005505$$

When X_1 — Plant height was dropped, the following model and information were obtained

$Y = 0.003651 + 0.0003079\ X_2 - 0.001730\ X_3 + 0.005289\ X_4$ (9)

Sum-of-squares (RSS) = 0.003956; SD of residuals = 0.01572; R squared = 0.1629;

F = 1.0382; The P value is 0.4024, VIF < 5 in all cases, n=20; K-l= 3;

$F_{(n_1,n_2,\infty)} = F_{(3,16,0.05)} = 3.24$

From the model above there was an increase of 0.0003079 and 0.005289 in the yield when each of the independent variable X_2 and X_4 was increased for every one unit increase respectively. However there was a decrease of about 0.001730 in the yield for every unit increase of the number of tillers. The intercept was at Y = 0.003651, the R^2- value of about 16% implied that about 16% of the total variation could be explained due to regression while the remaining percentage was unexplained. This shows that the model lacks goodness of fit. This is a poor fit, the F-calculated and P-value of 1.0382 and 0.4024 are indications of inadequacy and invalidity of the model. It will observe that F-calculated is less than the F-table value of 3.24 while the P-value is greater than 0.05 level of significance; the small value of R^2 and VIF proved the non-existence of multicollinearity.

$$PC_2 = \frac{RSS\{n+(K-1)\}}{\{n-(K-1)\}} = \frac{0.003956x23}{17} = 0.005352$$

When X_2 — number of leaves was dropped in the regression the following model and information were obtained

$Y = -0.07708 + 0.02880\ X_1 + 0.0004801\ X_3 + 0.004706\ X_4$ (10)

Sum-of-squares (RSS) = 0.003783; SD of residuals = 0.01538; R squared = 0.1995;

F = 1.3291; The P value is 0.2998; VIF < 5 in all cases, n = 20; K – l = 3;

$F_{(n_1,n_2,\infty)} = F_{(3,16,0.05)} = 3.24$

From the model, the intercept on y-axis is at –0.07708. There was an increase in the yield of 0.0288; 0.0004801 and 0.004706 respectively when each of the independent variables, plant height, number of tillers and the leafs square area was increased by one unit in that order. The R^2 — value of about 20% showed about 20% of the variation could be explained due to regression while the remaining percentage was unexplained. The F-value and the value calculated also confirmed the invalidity of the model because; the F-calculated of 1.3291 is less than the table value of 3.24 while the P-value calculated of 0.2998 is greater than 0.05 level of significance; The smaller value of R^2 and VIF which is less than in all cases area indications of no multicollinearity {R^2<0.75 and VIF<5}.

$$PC_3 = \frac{RSS\{n+(K-1)\}}{\{n^*-(K-1)\}} = \frac{0.003783x23}{17} = 0.005118$$

When X_3 — number of tillers was dropped, the following model and information were obtained

$$Y = -0.09519 + 0.03125\ X_1 + 0.0002645\ X_2 + 0.004795\ X_4 \quad (11)$$

Sum-of-squares (RSS) = 0.003689; SD of residuals= 0.01518;

R squared = 0.2 195; F = 1.4997; The P value is 0.2526, VIF < 5 in all cases, n = 20; K – 1 = 3; $F_{(n_1,n_2,\infty)} = F_{(3,16,0.05)} = 3.24$

From the model above, the intercept on the y-axis is at –0.09519. There was an increase in the yield of about 0.03125, 0.002645 and 0.004795 respectively when each of the independent variables above was plant height, number of leaves and leafs square area increase by one unit in that order. Because the F-calculated is less than 3.24 and the P-value of 0.2526 is greater than 0.05 level of significance, then it is evident that the model is invalid. However there was an improvement in the explained variation over the previous result when the leaf square area was dropped. The result showed that about 22% of the variation was explained or due to regression while the remaining percentage was unexplained. This also show that the model lack goodness of fit. The smaller value of R^2 =0.2195 and VIF less than 5 in all cases proved th non-existence of multicollinearity.

$$PC_4 = \frac{RSS\{n+(K-1)\}}{\{n-(K-1)\}} = \frac{0.003689x23}{17} = 0.004991$$

When X_4 — leafs square area was dropped in the regression the following model of y versus X_1, X_2, and X_3 were obtained with the following information

$$Y = -0.08781 + 0.03075\ X_1 + 0.0001482\ X_2 + 0.002084\ X_3 \quad (12)$$

Sum-of-squares (RSS) = 0.004304; SD of residuals = 0.01640; R squared = 0.0892;

F= 0.5225; The P value is 0.6729; VIF<5 in all cases, n=20; K-l= 3;

$F_{(n_1,n_2,\infty)} = F_{3,16,0.05)} = 3.24$

From the model above, the intercept on the y-axis is at –0.08781. For every unit increase of plant height, number of leaves and number of tillers there was an increase of 0.03075, 0.001482 and 0.002084 in the yield respectively. The R^2 value indicated that about 9% of the variation was explained, that is due to regression while the remaining percentage was unexplained. This is a poor model, F-value calculated and the P-value of 0.5225 and 0.6729 respectively proved

that the model is invalid. Although the VIF which is less than five in all cases and the smaller value of R^2 implied there is no multicollinearity but the model lacks the goodness of fit.

$$PC_5 = \frac{RSS\{n+(K-1)\}}{\{n-(K-1)\}} = \frac{0.004304x23}{17} = 0.005823$$

Using the Rizzi Laura prediction criterium (PC) arranging the $PC_{i's}$ in ascending order gives $PC_4 < PC_3 < PC_2 < PC_1 < PC_5$ where $PC_4 = 0.004991$, $PC_3 = 0.005118$, $PC_2 = 0.005352$, $PC_i = 0.005505$ and $PC_5 = 0.005823$. From the PC_4 is the least prediction criterium, the model having this PC_4 is the best among several other that has the best goodness of fit for this variety. This is when the number of tillers was dropped in the regression. If the exclusion of number of tiller leads to having the best model, then the number of tillers is an irrelevant plant attribute for the variety (Dauro).

Conclusion

The following model was arrived at as being the best model that fairly satisfied the necessary and required conditions used to assess the goodness of fit of a model. The best model that fits the data set well was obtained when the number of tiller's was excluded as an independent variable or covariate in the regression analysis of Dauro variety. The next best model was obtained when the number of leaves was excluded in the regression analysis. Finally, the worst model was obtained when leafs square area was dropped in the analysis.

REFERENCES

Baker, E.F.I (1970): KenafandRoselle in Western Nigeria, Word crop Nov-Dec 1970 Pp380-386

Busari, A.F; Abubakar .U. and Cole.A.T (2010): Constructing the Best Regression model for maiwa variety. Pakistan Journal of Nutrition 9(4):380-386

Garder B.R and T.C. Tucker (1976): Nitrogen effects on Cotton: Vegetative and fruiting characteristics Soil/Science America Proc. vol 31. No. 6 Nov-Dec pp780-791-56

Ogunremi E.A (1970): Relationship between yields and some Agronomic Characters of sugar cane in Southern Nigeria Nig. Agric. Journal pp 1-8

Rizzi Laura (2008): Specification error multicollinearity and Qualitative covariate. www.site net Nov 19, 2008

Sarah P. Otto and Troy day (2007): A Biologists Guide to Mathematical Modeling in Ecology and Evolution. Princeton University; New Jersey 08540; PP 17-19; 567

APPENDIX 1

Multiple Regression Results; what equation fits the data the best ?
$Y = -0.09209 + 0.03008 X_1 + 0.0003453 X_2 - 0.001077 X_3 + 0.005237 X_4$
How good is the fit? R squared = 22.33%.
This is the percent of the variance in A: Yield(Y) explained by the model.
The P value is 0.4018, considered not significant.
The P value answers this question: If there were no linear relationship among the variables, what is the chance that R squared would be that high (or higher) by chance?
Since P is high, the rest of the results will be of little interest.
Sum-of-squares 0.003670; SD of residuals 0.01564; R squared 0.2233;
Adjusted R squared 0.0162; Multiple R 0.4726; F 1.0784
Is multicollinearity a problem ?

Variable	VIF	R2 with other X
B: Plant height (Xl)	1.07	0.0682
C: No. Leaves(X2)	1.58	0.3682

D: No. tillers(X3)	2.03	0.5074
E: Leaf sq.area(X4)	1.35	0.2574

Each R squared quantifies how well that X variable is predicted from the other X variables (ignoring Y).
VIF is calculated from R squared.
All R squared values are low (<0.75). The X variables are independent of each other.
Multicollinearity is not a problem.****
Multiple Regression Results; what equation fits the data the best ?
$Y = 0.003651 + 0.0003079\ X_2 - 0.001730X_3 + 0.005289\ X_4$
How good is the fit? R squared = 16.29%.
This is the percent of the variance in A: Yield(Y) explained by the model.
The P value is 0.4024, considered not significant.
The P value answers this question:
If there were no linear relationship among the variables, what is the chance that R squared would be that high (or higher) by chance?
Since P is high, the rest of the results will be of little interest.
Sum-of-squares 0.003956; SD of residuals 0.01572; R squared 0.1629
Adjusted R squared 0.0060; Multiple R 0.4037; F 1.0382
Is multicollinearity a problem?

Variable	VIF	R2 with other X
C:No. Leaves(X2)	1.58	0.3653
D:No. tillers(X3)	1.98	0.4955
E:Leafsq.ar(X4)	1.35	0.2573

Each R squared quantifies how well that X variable is predicted from the other X variables (ignoring Y).
VIF is calculated from R squared.
All R squared values are low (<0.75). The X variables are independent of each other.
Multicollinearity is not a problem.***
Multiple Regression Results; what equation fits the data the best?
$Y = -0.07708 + 0.02880\ X_2 + 0.0004801\ X_3 + 0.004706\ X_4$
How good is the fit? R squared = 19.95%.
This is the percent of the variance in A:Yield(Y) explained by the model.
The P value is 0.2998, considered not significant.
The P value answers this question:
If there were no linear relationship among the variables, what is the chance that R squared would be that high (or higher) by chance?
Since P is high, the rest of the results will be of little interest.
Sum-of-squares 0.003783; SD of residuals 0.01538; R squared 0.1995
Adjusted R squared 0.0494; Multiple R 0.4466; F 1.3291
Is multicollinearity a problem?

Variable	VIF	R2 with other X
B: Plant height(Xl)	1.07	0.0638
D:No. tillers(X3)	1.34	0.2556
E:Leafsq.are(X4)	1.27	0.2118

Each R squared quantifies how well that X variable is predicted from the other X variables (ignoring Y).
VIF is calculated from R squared.
All R squared values are low (<0.75). The X variables are independent of each other.
Multicollinearity is not a problem.***
Multiple Regression Results; what equation fits the data the best?
$Y = -0.09519 + 0.03125X_1 + 0.0002645X_2 + 0.004795X_4$
How good is the fit? R squared = 21.95%.
This is the percent of the variance in A:Yield(Y) explained by the model.
The P value is 0.2526, considered not significant.
The P value answers this question:
If there were no linear relationship among the variables, what is the chance that R squared would be that high (or higher) by chance?
Since P is high, the rest of the results will be of little interest.

Sum-of-squares 0.003689; SD of residuals 0.01518; R squared 0.2195
Adjusted R squared 0.0731; Multiple R 0.4685; F 1.4997
Is multicollinearity a problem?

Variable	VIF	R2 with other X
B: Plant height(Xl)	1.05	0.0457
C:No. Leaves(X2)	1.05	0.0454
E:Leafsq.are(X4)	1.01	0.0124

Each R squared quantifies how well that X variable is predicted from the other X variables (ignoring Y).
VIF is calculated from R squared.
All R squared values are low (<0.75). The X variables are independent of each other.
Multicollinearity is not a problem.**
Multiple Regression Results; what equation fits the data the best?
$Y = -0.08781 + 0.03075X_1 + 0.0001482X_2 + 0.002084X_3$
How good is the fit? ; R squared = 8.92%.
This is the percent of the variance in A:Yield(Y) explained by the model.
The P value is 0.6729, considered not significant.
The P value answers this question:
If there were no linear relationship among the variables, what is the chance that R squared would be that high (or higher) by chance?
Since P is high, the rest of the results will be of little interest.
Sum-of-squares 0.004304; SD of residuals 0.01640; R squared 0.0892
Adjusted R squared –0.0815; Multiple R 0.2987; F 0.5225
Is multicollinearity a problem?

Variable	VIF	R2 with other X
B: Plant height(Xl)	1.07	0.0680
C:IMo. Leaves(X2)	1.49	0.3294
D:No. tillers(XS)	1.53	0.3449

Each R squared quantifies how well that X variable is predicted from the other X variables (ignoring Y).
VIF is calculated from R squared.
All R squared values are low (<0.75). The X variables are independent of each other.
Multicollinearity is not a problem.

APPENDIX-2

Table-1 : Data on Measurement of Dauro Cereal Crop

S/NO	YIELD(Y)	PLANT HIEGHT (X_1)	NO. OF LEAVES (X_2)	NO. OF TILLERS (X_3)	LEAFS SQUARE AREA (X_4)
1	0.0388	3.14	27.6	2.0	0.630
2	0.0207	3.22	39.8	3.0	0.989
3	0.0550	2.94	43.8	5.0	5.370
4	0.0226	3.12	43.8	5.0	5.370
5	0.0187	3.10	56.6	4.0	1.536
6	0.0085	3.30	38.6	2.0	1.249
7	0.0043	2.96	48.2	3.0	1.631
8	0.0409	3.12	51.0	4.0	1.476
9	0.0505	3.06	60.2	5.0	1.547
10	0.0275	3.14	60.2	4.0	1.547
11	0.0353	3.12	49.4	2.0	1.348
12	0.0161	2.84	39.0	3.0	1.196
13	0.0138	3.02	42.8	4.0	1.494
14	0.0104	3.18	44.2	5.0	1.096
15	0.0165	3.02	52.2	6.8	1.874
16	0.0035	3.02	33.8	2.0	0.905
17	0.0105	3.00	45.2	3.0	1.076
18	0.0040	3.00	50.4	4.0	1.704
19	0.0070	2.74	60.6	5.0	1.486
20	0.0072	2.88	41.6	3.6	1.121