International Journal in IT & Engineering (IJITE)
Volume 5 Issue 12, December 2017 ISSN: 2321-1776 Impact Factor: 6.341
Journal Homepage: http://ijmr.net.in, Email: irjmss@gmail.com
Double-Blind Peer Reviewed Refereed Open Access International Journal

DATA ANALYTICS AND PREDICTIONS IN CROP YIELDING

Dr.R.Jamuna**

****Professor, Department of Computer Science S.R.College, Bharathidasan University, Trichy.**

ABSTRACT

Regression technique can be best suited for predications in agriculture. In this paper regression analysis models the relationship between factors like plant height and tiller number which are independent variables with the yields of a crop (like rice plant) which is a dependent variable that we want to predict. In data mining independent variables are attributes already known and response variables are what we want to predict. Data analytics here involves regression analysis with more than one independent variable which is called *multiple regression analysis.* When all independent variable are assumed to affect the dependent variable in a linear proportion and independently of one another, the procedure is called *multiple linear regression analysis*. The simple liner regression and correlation analysis has one major limitation. That is applicable only to cases with one independent variable. There is a corresponding increase in need for use of regression procedures that can simultaneously handle several independent variables. Thus, the combined linear effects of plan height and tiller number with the variation in yield can be predicted with the computation of F values from Test of significance. The idea can be extended for many software based predictions where the computational steps of SSR and SSE can be derived from the outputs of software coding with saving of time and accuracy when number of samples increase with more independent variables in Big data analytics

I INTRODUCTION

Regression technique can be best suited for predications in agriculture. In this paper regression analysis models the relationship between factors like plant height and tiller number which are independent variables with the yields of a crop (like rice plant) which is a dependent variable. In data mining independent variables are attributes already known and response variables are what we want to predict. Data analytics here involves regression analysis with more than one independent variable which is called *multiple regression analysis.* When all independent variable are assumed to affect the dependent variable in a linear proportion and independently of one another, the procedure is called *multiple linear regression analysis*. The simple liner regression and correlation analysis has one major limitation. That is, that it is applicable only to cases with one independent variable. There is a corresponding increase in need for use of regression procedures that can simultaneously handle several independent variables. A multiple linear regression is said to be operating if the relationship of the dependent variable *Y* to the *k* independent variables X_1, X_2, X_k can be expressed as[1] $Y=\alpha+\beta+ \beta_1X_1+ \beta_2X_2 \ ...+ \beta_kX_k$. The data required for the application of the multiple linear Regression analysis involving *k* independent variables are the (*n*) (*k*+1) observation described here:

OBSERVATION VALUE

Observation number	Y	$\bar{X}_1$	$\bar{X}_2$	$\bar{X}3$	…..	$\bar{X}_k$
1	Y_1	Y_{11}	Y_{21}	Y_{31}	….	Y_{k1}
2	Y_2	Y_{12}	Y_{22}	$Y_{32}2$	….	Y_{k2}
3	Y_3	Y_{13}	Y_{23}	Y_{33}	….	Y_{k3}
..	..	..	..	..	..	..
n	Y_n	X_{1n}	X_{2n}	X_{3n}	….	X_{kn}

The (k+1) variables Y, $\bar{X}_1$, $\bar{X}_2$,… $\bar{X}_k$ must be measured simultaneously for each of the n units of observations (i.e., experimental unit or sampling unit) in addition ,there must be enough observation to make n greater than(k+1)..The multiple linear regression procedure involves the estimations and test of significations of the (k+1) parameters of the multiple linear regression equation. We illustrate the procedure for a case where k =2 ,using the **data on grain yield (Y), plan height (X_1), and tiller number (X_2)** in table 9.6 with k=2, the multiple linear regression equation is expressed as:

$$Y=\alpha+\beta_1 X_1+\beta_2 X_2$$

II COMPUTATIONAL METHOD FOR DATA ANALYTICS

Compute the mean and the corrected sum of squares for each of the (k+1) variables Y X_1, X_2….X_K, and the corrected sums of cross production for all possible pair-combinations of the (k+1) variables, [2]

Corrected sum of squares and cross products						
Variables	**Mean**	X_1	X_2	…….	X_k	y
X_1	$\bar{X}_1$	$\sum x_1^2$	$\sum x_1 x_2$	….	$\sum x_1 X_k$	$\sum x_1 y$
X_2	$\bar{X}_2$		$\sum x_2^2$	….	$\sum x_2 X_k$	$\sum x_2 y$
…	….	….	….	…	…	…
…	..	….	….	….	….	….
X_k	$\bar{X}_k$				$\sum x_k^2$	$\sum x_k y$
Y	$\bar{Y}$					$\sum x^2$

Table [1.1] COMPUTATION OF A MULTIPLE LINEAR REGRESSION EQUATION

Relating plant height (X1) and tiller number (X2) to yield (Y).(8 plant varieties like rice)

Variety number	Grain yield, kg/ha(Y)	Plant height, cm (X_1)	Tiller no./hill (X_2)
1	5,755	110.5	14.5
2	5,939	105.4	16.0
3	6,010	118.1	14,6
4	6,545	104.5	18.2
5	6,730	93.6	15.4
6	6,750	84.1	17.6
7	6,899	77.8	17.9
8	7,862	75.6	19.4
Mean	6,561	96.2	16.7
$\sum x_1^2$ =1,753.72	$\sum x_2^2$ = 23.22		$\sum x_1 y$ = -65,194
$\sum x_2 y$= 7,210	$\sum x_1 x_2$= -156.65		$\sum y^2$= 3,211,504

III PREDICTIONS FROM THE DATA ANALYTICS

Linear function of *k* independently variables to the variations in *Y*. it is usually expressed in percentage .Its square root (i.e. R^2) is referred to as the *multiple correlation coefficient.* For our example, the values of *SSR*, R^2, and *SSE* are computed as: [3]

$$SSR = b_1 \sum x_1 y + b_2 \sum x_2 y$$

$$= (-23.75)(-645,194)+(150.27)(7,210)$$

$$= 2,631,804$$

$$SSE = \sum y^2 - SSR = 3,211,504 - 2,631,804$$

$$= 579,700$$

$$R^2: \quad = \frac{SSR}{} = 2,631.804 \quad = \quad 82$$

IV. RESULTS AND CONCLUSIONS

Thus, 82% of the total variations in the yield of eight plant varieties can be accounted for by a linear function, involving plant height and tiller number, as expressed in step 3.

Step 5: Test the signification of R^2:

Computer the F value as:

$$F = \frac{SSR/k}{SSE/(n-k-1)} \qquad F = \frac{2{,}631{,}804/2}{579{,}900/(8-2-1)} = 11.35$$

Compare the computed F value to the tabular F values with f = K and f = (n-k-1) degrees of freedom. The coefficient of determination R is said to be significant.[4] (Significantly different from zero) if the computed F value is greater than the corresponding tabular F value at the prescribed level of significance. For our example, the tabular F values with f = 2^{nd} f = 5 degrees of freedom are 5.79 at the 5% level of significance and 13.27 at the 1% level. Because the computer F value is larger than the corresponding tabular F value at the 5% level of significance, but smaller than the tabular F value at the 1% level, the estimated multiple liner Regression Y = 6,336 – 23.75 X + 150.27X is significant. **Thus, the combined linear effects of plan height and tiller number contribute significantly to the variation in yield.** The idea can be extended for large data by software based predictions .The computational steps of SSR and SSE can be derived from the outputs of software coding with saving of time and accuracy when number of samples increase with more independent variables in big data analytics .

V. REFERENCES

[1] Statistical Package for the Social Sciences, 2nd ed. USA: McGraw-Hall, 1975 pp A.J. Barr et al. Analysis System User's Guide. USA: SAS Institute, 1979. pp. 237-263

[2] John P. Hoffmann, ", USA. Linear Regression Analysis: Applications And Assumptions Second Edition", 2010

[3] Dan Campbell, Sherlock Campbell, "Introduction to Regression and Data Analysis", October 28, 2008.

[4] Barr et al., 1972; Nie et al., 1975: and Dixon, 1975