Available online at www.isas.org.in/jisas
Journal of the Indian Society of Agricultural Statistics 73(3) 2019 207–210

Use of Multivariate Multiple Regression Model to Predict Wet-season Rice Yield in Lower Gangetic Region of West Bengal, India

Ria Biswas and Banjul Bhattacharyya
Bidhan Chandra Krishi Viswavidyalaya, Nadia

Received 25 September 2017; Revised 26 June 2019; Accepted 30 July 2019

SUMMARY

Multivariate multiple regression considers the linear relationship between one or more dependent variables and more than one independent variables. To predict rice yield, the transplanting date (here used as dummy variable (0, 1)) along with other weather variables were used as independent variables. Wet-season rice yield and biomass were used as dependent variable. Multivariate multiple regression method was applied to predict the crop yield and biomass simultaneously. It was observed that the model performance is quite. It can be concluded that such model using Multivariate multiple regression can be used for total yield and biomass prediction at district level.

Keywords: Multivariate multiple regression, Wet-season rice, Yield forecasting.

1. INTRODUCTION

Rice (*Oryza sativa*) is the most important staple food and predominant crop in the South East Asian countries (IRRI, 2006). In India, rice is grown in 43.8 million hectares land (CRRI, 2011). Crop production is highly dependent on weather parameters. The changes of meteorological parameters such as, variability of rainfall and monsoon shifting along with temperature rise influence crop yields and productivity. Fluctuations of weather parameters create various problems in rice production (Anandakumar *et al.* 2008, Banerjee, 2008, Banerjee *et al.* 2014). Yield variation due to climatic uncertainty leads to food insecurity and fluctuation of its price. Forecasts of crop yield can provide important information about commodity markets and are frequently used by growers, industry and government to make decisions in advance (Vogel and Bange, 1999).

To fix the procurement price and to regulate the Government distribution yield forecasting is necessary. Since 1920 crop-weather relationship model for forecasting yield was used popularly (Fisher, 1924). To establish crop-weather relationship and predict the yield regression model is very useful tool (Rahman *et al.* 2005, Biswas *et al.* 2015). But, if more than one dependent variable present one must perform regression analysis separately for two dependent variables along with other independent variables. This process is time consuming and laborious too. A simple remedy to overcome this situation is to use multivariate multiple regression analysis where more than one variable is considered as dependent variable. Considering the background, the present research work aims to predict the yield and biomass of wet-season rice simultaneously.

2. MATERIALS AND METHODS

2.1 Study area

The study was conducted for Nadia District, West Bengal, Eastern India which falls under New Alluvial Agro-climatic Zone of Gangetic West Bengal. The climate of this said zone is sub-tropical with mean annual rainfall varying from 1400 to 1700 mm. More than 70 % of total annual rainfall is received from South-West monsoon during the months of June to September. The average temperature varies from 15.6 to 35ºC. The main crop of this district is wet-season rice. Water requirement of the said crop is met mainly from the

Corresponding author: Ria Biswas
E-mail address: ria.biswas6@gmail.com

monsoonal rainfall. Due to erratic and irregular pattern of rainfall, farmers have to depend upon irrigation.

2.2 Secondary data collection

Wet-season rice yield (Satabdi variety) and biomass data with different transplanting dates were collected from Annual Progress Report of AICRP on Agro-meteorology of Mohonpur Centre, Directorate of Research, Bidhan Chandra Krishi Viswavidyalaya (State Agriculture University). Meteorological observatory of Kalyani (22.67°N, 88.20°E and 7.8 m above mean sea level), Nadia District provided weather data, namely, maximum temperature, minimum temperature, vapor pressure deficit, wind speed, rainfall, relative humidity, bright sun shine hour, evaporation and number of rainy days.

In the present study to include the effect of date of transplanting on rice yield and biomass, the entire transplanting season has been classified into two categories. First half transplanting (early transplanting date to normal transplanting date) scored as zero (0) and second half transplanting (i.e. late transplanting) scored as one (1) to treat them as dummy variable. The weather variables and the transplanting dates are defined in Table 1.

Table 1. Weather parameters used in models with their notations

Sl No.	Weather parameters	Notations
1	Maximum temperature	Z1
2	Minimum temperature	Z2
3	Vapor pressure deficit	Z3
4	Wind speed	Z4
5	Rainfall	Z5
6	Relative humidity	Z6
7	Bright sun shine hour	Z7
8	Evaporation	Z8
9	Number of rainy days	Z9
10	Transplanting date	Z10

For development of statistical model, the weather data along with wet-season rice yield and biomass data were collected for twelve years (2002-2013) i.e 24 number of data sets. The output of the developed model was compared with actual data set for the year 2014 and 2015.

2.3 Methodology

Multivariate multiple regression consider the linear relationship between two or more *y*'s (the *dependent* or *response* variables) and one or more *x's* (the *independent* or *predictor* variables). Linear model will be used to relate the *y*'s to the *x's* and will be concerned with estimation and testing of the parameters in the model. One aspect of interest will be choosing which variables to include in the model if this is not already known (Rencher, 2002).

Let us take this example where we want to know simultaneously the yield (Y_1) of wet-season rice grain yield and biomass yield (Y_2) from the same set of explanatory variables like X_1, X_2, X_3,……., X_{10}.

So using the technique of multivariate regression we cane frame two regression equations:

$$Y_1 = \alpha_1 + \sum_{i=1}^{10} \beta_{i1} X_i + \varepsilon_1 \quad (1)$$

and

$$Y_2 = \alpha_2 + \sum_{i=1}^{10} \beta_{i2} X_i + \varepsilon_2 \quad (2)$$

By shifting the origin of the variables to their respective means one can avoid yhe intercept terms (α) in the regression without loosing any generality and in matrix notation these two equations can be written as

$$\underset{\sim}{Y}_1 = \underset{\sim}{X}\underset{\sim}{\beta}_1 + \underset{\sim}{\varepsilon}_1 \quad (3)$$

and

$\underset{\sim}{Y}_2 = \underset{\sim}{X}\underset{\sim}{\beta}_2 + \underset{\sim}{\varepsilon}_2$ where $\underset{\sim}{Y}_1$, $\underset{\sim}{Y}_2$, $\underset{\sim}{\varepsilon}_1$, $\underset{\sim}{\varepsilon}_2$ are $n \times 1$ matrices; (4)

$\underset{\sim}{\beta}_1$ and $\underset{\sim}{\beta}_2$ are 10×1 matrices and $\underset{\sim}{X}$ is a $n \times 10$ matrix.

In multivariate regression analysis the above systems are combined together to get

$$[\underset{\sim}{Y}_1, \underset{\sim}{Y}_2] = \underset{\sim}{X}[\underset{\sim}{\beta}_1, \underset{\sim}{\beta}_2] + [\underset{\sim}{\varepsilon}_1, \underset{\sim}{\varepsilon}_2] \quad (5)$$

$$\text{Or, } \underset{\sim}{Y} = \underset{\sim}{X}\underset{\sim}{\beta} + \underset{\sim}{\varepsilon}, \quad (6)$$

where, $\underset{\sim}{Y}$ is $n \times 2$ matrix of the dependent variables, $\underset{\sim}{\beta}$ is 10×2 matrix of parameters, $\underset{\sim}{X}$ is $n \times 10$ matrix of explanatory variables and $\underset{\sim}{\varepsilon}$ is $n \times 2$ matrix of error terms.

The least squares estimates for b's is given by

$$\hat{\underset{\sim}{\beta}} = [\underset{\sim}{X}' \underset{\sim}{X}]^{-1}[\underset{\sim}{X}' \underset{\sim}{Y}] \quad (7)$$

where

$$[\underset{\sim}{X}' \underset{\sim}{X}]^{-1} = \left[\begin{bmatrix} X & 0 \\ 0 & X \end{bmatrix}^1 \begin{bmatrix} X & 0 \\ 0 & X \end{bmatrix}\right]^{-1} = \begin{bmatrix} X'X & 0 \\ 0 & X'X' \end{bmatrix}^{-1}$$

$$= \begin{bmatrix} [X'X]^{-1} & 0 \\ 0 & [X'X]^{-1} \end{bmatrix}$$

and $[X' Y] = \begin{bmatrix} X & 0 \\ 0 & X \end{bmatrix}^1 \begin{bmatrix} Y_1 \\ Y_2 \end{bmatrix} = \begin{bmatrix} X'Y_1 \\ X'Y_2 \end{bmatrix}$

$$\hat{\beta} = \begin{bmatrix} \hat{\beta}_1 \\ \hat{\beta}_2 \end{bmatrix} = \begin{bmatrix} [X'X]^{-1} & 0 \\ 0 & [X'X]^{-1} \end{bmatrix} \begin{bmatrix} X'Y_1 \\ X'Y_2 \end{bmatrix}$$

hence the least square estimates for

$$\hat{\beta}_1 = [X'X]^{-1}[X'Y_1] \text{ and} \quad (8)$$

$$\hat{\beta}_2 = [X'X]^{-1}[X'Y_2] \quad (9)$$

The above estimates are the estimates of the coefficients of multiple linear regression equations of Y_1 and Y_2 separately. Thus the regression parameters of multivariate regression analysis can be worked out following the estimation of regression parameters in multiple regression equations separately on each of the dependent variable.

3. RESULTS AND DISCUSSION

3.1 Development of Multivariate multiple regression equation

The steps to formulate the Multivariate multiple regression equation are discussed in this section. The detailed calculation process is as follows:

$$\hat{\mu} = \begin{bmatrix} \bar{y} \\ \bar{z} \end{bmatrix} = \begin{bmatrix} 3818.635 \\ 8733.079 \\ \cdots\cdots\cdots \\ 33.327 \\ \vdots \\ 0.5 \end{bmatrix}$$

and

$$S = \begin{bmatrix} S_{YY} & \vdots & S_{YZ} \\ \cdots\cdots\cdots\cdots & & \\ S_{ZY} & \vdots & S_{ZZ} \end{bmatrix}$$

where,

$$S_{YY} = \begin{bmatrix} 738321.8721 & -838686.8451 \\ -838686.8451 & 6937869.044 \end{bmatrix}$$

$$S_{YZ} = \begin{bmatrix} 439.16 & 104.00 & 7.57 & -34.49 & -23793.03 \\ 571.07 & 108.95 & 56.84 & 717.21 & -218210.30 \\ -52.75 & 29.65 & 70.59 & -1250.35 & -200.86 \\ -2198.77 & 759.46 & 504.88 & -6778.88 & -292.15 \end{bmatrix}$$

$$S_{ZY} = \begin{bmatrix} 439.16 & 571.07 \\ 104.00 & 108.95 \\ 7.57 & 56.84 \\ -34.49 & 717.213 \\ -23793.03 & -218210.30 \\ -52.75 & -2198.77 \\ 29.65 & 759.46 \\ 70.59 & 504.88 \\ -1250.35 & -6778.88 \\ -200.86 & -292.15 \end{bmatrix}$$

$$S_{ZZ} = \begin{bmatrix} 1.45 & 0.08 & 0.00 & 0.37 & -84.63 & -1.10 & 0.55 & 0.40 & -5.31 & -0.45 \\ 0.08 & 0.31 & 0.01 & -0.27 & -34.85 & -0.74 & 0.00 & 0.01 & -1.39 & -0.04 \\ 0.00 & 0.01 & 0.00 & -0.01 & -3.47 & -0.02 & 0.008 & 0.00 & -0.05 & -0.00 \\ 0.37 & -0.27 & -0.01 & 0.73 & -0.83 & 0.01 & 0.19 & 0.24 & -1.44 & -0.16 \\ -84.63 & -34.85 & -3.47 & -0.83 & 27317.18 & 124.60 & -53.06 & -25.01 & 410.15 & 19.05 \\ -1.10 & -0.74 & -0.02 & 0.01 & 124.60 & 3.68 & -0.39 & -0.49 & 8.84 & 0.61 \\ 0.55 & 0.00 & 0.00 & 0.19 & -53.06 & -0.39 & 0.48 & 0.21 & -2.27 & -0.13 \\ 0.40 & 0.01 & 0.00 & 0.24 & -25.01 & -0.49 & 0.21 & 0.21 & -2.38 & -0.16 \\ -5.31 & -1.39 & -0.05 & -1.44 & 410.15 & 8.84 & -2.27 & -2.38 & 38.25 & 2.23 \\ -0.45 & -0.04 & -0.00 & -0.16 & 19.05 & 0.61 & -0.13 & -0.16 & 2.23 & 0.26 \end{bmatrix}$$

Assuming normality, the estimated regression function is,

$$\hat{\beta}_0 + \hat{\beta}_z = \bar{y} + S_{yz} S_{zz}^{-1} (z - \bar{z}) \quad (10)$$

$$= \begin{bmatrix} 3818.635417 \\ 8733.079861 \end{bmatrix} + \begin{bmatrix} 225.09 & 1811.07 & -2003.59 & 95.89 & -0.58 & 673.49 & -355.77 & 508.67 & 18.33 & -1575.30 \\ -1164.58 & -4412.69 & 26502.63 & 851.32 & -4.48 & -1425.94 & 1768.23 & -3767.40 & -241.69 & 1098.62 \end{bmatrix} \begin{bmatrix} z_1 - 33.32 \\ z_2 - 26.18 \\ z_3 - 3.56 \\ z_4 - 1.12 \\ z_5 - 376.54 \\ z_6 - 94.86 \\ z_7 - 5.01 \\ z_8 - 2.65 \\ z_9 - 27.45 \\ z_{10} - 0.5 \end{bmatrix}$$

Thus the best predictor of Y_1 is,

3818.63 + 2226.09(Z_1-33.32) + ……. – 1676.30(Z_{10} – 0.5)

= -107014 + 2226.09Z_1 + 1811.07 Z_2 - 2003.60 Z_3 + 96.89 Z_4 - 0.68 Z_5 + 673.49 Z_6 -

366.77Z_7+608.67Z_8+18.33Z_9-1676.31Z_{10} (11)

Similarly, the best predictor of Y_2 is,

8733.07 – 1164.68(Z_1-33.32) + ……+ 1098.60(Z_{10} – 0.5)

= 211726 - 1164.68 Z_1 - 4412.69 Z_2 + 26602.63 Z_3 + 861.3229 Z_4 - 4.48146 Z_5 - 1426.96

Z_6 + 1768.236 Z_7 - 3767.4 Z_8 - 241.694 Z_9 + 1098.621 Z_{10} (12)

The maximum likelihood estimate of the expected squared errors and cross-products matrix ∑YY. Z is given by,

$$\left(\frac{n-1}{n}\right)\left(S_{YY}-S_{YZ}S_{ZZ}^{-1}S_{ZY}\right) \quad (13)$$

$$=\left(\frac{24-1}{24}\right)\begin{bmatrix}172260.12 & -89569.94\\ -89569.94 & 1095125.51\end{bmatrix}$$

$$=\begin{bmatrix}165082.61 & -85837.86\\ -85837.86 & 1049495.28\end{bmatrix}$$

The first estimated regression function, -107014 + 2226.09Z_1 + 1811.07 Z_2 - 2003.60 Z_3 + 96.89 Z_4 - 0.68 Z_5 + 673.49 Z_6 - 366.77 Z_7 + 608.67 Z_8 + 18.33 Z_9 - 1676.31 Z_{10}, and the associated mean square error, 165082.61 for the single-response case.

Similarly, the second estimated regression function, 211726 - 1164.68 Z_1 - 4412.69 Z_2 + 26602.63 Z_3 + 861.3229 Z_4 - 4.48146 Z_5 - 1426.96 Z_6 + 1768.236 Z_7 - 3767.4 Z_8 - 241.694 Z_9 + 1098.621 Z_{10} and the associated mean square error, 1049496.28 for the single-response case.

It is observed that the data enable to predict the first response, Y_1, with smaller error than the second response, Y_2. The negative covariance -85837.86 indicated that over-prediction of rice grain yield tends to be accompanied by under-prediction of rice biomass yield.

3.2 Performance of the Model

The weather data along with transplanting dates of 2014 and 2015 were used to validate the yields. After putting the value of different explanatory variables (as indicated in the previous section) in equations 11 and equation 12 the forecasted yield values were obtained (Table 2).

4. CONCLUSION

The study reveals that the equations enable to predict the crop yield and biomass simultaneously and required less time for data analysis. In general, we performed regression analysis separately for two dependent variables along with other independent variables, but in multivariate multiple regression analysis more than one variable is considered as dependent variable. From the normal weather data of the district and date of onset of monsoon the yield of wet-season rice along with biomass can be predicted jointly from a single calculation which will be very much helpful for forecast planner.

REFERENCES

Anandakumar, S., Subramani,T. and Elango, L. (2008). Spatial Variation and Seasonal Behaviour of Rainfall Pattern in Lower Bhavani River Basin, Tamil Nadu, India. *The Eco-Scan,* **2(4)**, 17-24.

Banerjee S, (2008). Possible impact of climate change on rice production in the Gangetic West Bengal, India, In: Unkovich MJ (ed) Global Issues Paddock Action, Proceedings of the *14th Australian Agronomy Conference*, 21-25 September 2008, Adelaide, South Australia.

Banerjee, S., Das, S.,Mukherjee, A., Mukherjee, A., Saikia, B. (2014) Adaption strategies to combat climate change effect on rice and mustard in Eastern India, *Mitig Adapt Strateg Glob Change*, DOI 10.1007/s11027-014-9595-y.

Biswas, R., Bhattacharyya, B. and Banerjee, S. (2015). Predicting wet-season rice yield of Gangetic West Bengal through weather based regression model using dummy variable, *The Ecoscan*, **9(1&2)**, 37-41.

CRRI (2011), Vision 2030. Central Rice Research Institute. ICAR, India.

Fisher, R.A. (1924). The influence of rainfall on the yield of wheat at Rothamsted, *Roy. Sco. (London) Phil. Trans. Ser.* **B 213**, 89-142.

IRRI (2006), Bringing hope, improving lives: Strategic Plan 2007-2015, Manila 61 p.

Rahman, S.M., Md. Huq, M., Sumi, A., Mostafa, G.M., Azad, R.M. (2005). Statistical Analysis of Crop-Weather Regression Model for Forecasting Production Impact of Aus Rice in Bangladesh. *Int. J. Statist. Sci.*, **4**, 57-77.

Rencher, A.C. (2002). Methods of Multivariate Analysis, Second edition, *A John Wiley & Sons, Inc*. publication.

Vogel F, Bange G. (1999) Understanding crop statistics. (available at : www.usda.gov/nass/nassinfo/pub1554.htm.).

Table 2. Comparison between actual value and predicted value of rice yield and biomass through statistical model

Year	Date of transplanting	Yield (Y_1) (Kg ha^{-1})			Biomass (Y_2) (Kg ha^{-1})		
		Actual	Predicted	Accuracy %	Actual	Predicted	Accuracy %
2014	d_0(early to normal transplanting)	6944	7935.98	85.71	8975	7816.79	87.09
	d_1(late transplanting)	5802	6832.57	82.23	5950	7209.72	78.82
2015	d_0 (early to normal transplanting)	5932	7102.19	80.27	9650	8326.89	86.28
	d_1(late transplanting)	5235	6313.54	79.39	9825	8772.45	89.28