MAUSAM, **67**, 4 (October 2016), 913-918

633 : 551.509.314

Use of ordinal logistic regression in crop yield forecasting

VANDITA KUMARI, RANJANA AGRAWAL and AMRENDER KUMAR*

Indian Agricultural Statistics Research Institute, Library Avenue, Pusa, New Delhi – 110 012, India

***Agricultural Knowledge Management Unit (AKMU),*

Indian Agricultural Research Institute, Pusa, New Delhi – 110 012, India

(*Received 16 July 2013, Accepted 22 June 2016*)

e mail : vandita.iasri@gmail.com

सार – उत्तर प्रदेश के कानपुर जिले के गेहूँ की फसल के लिए उपज पैदावार पूर्वानुमान में क्रमसूचक लॉजिस्टिक समाश्रयण और विभेदी विश्लेषण के उत्पादन की तुलना की गई। फसल वर्षों को डीट्रेंडिड पैदावार के आधार पर दो अथवा तीन समूहों में विभाजित किया गया। परवर्ती वर्षों के आंकड़ों का उपयोग करते हुए समाश्रयण और मान्यकृत वर्षों के रूप में क्रमसूचक लॉजिस्टिक समाश्रयण के माध्यम से प्राप्त की गई। प्राथमिकताओं का उपयोग करके फसल पैदावार पूर्वानुमान मॉडल विकसित किए गए। विभेदी प्रकार्य के लिए दो प्रकार के मॉडल विकसित किए गए, एक में स्कोर का उपयोग किया गया और दूसरे में उत्तर प्रायिकताओं का उपयोग किया गया। विभिन्न सप्ताहों में प्राप्त किए गए मॉडल के निष्पादन की तुलना Adj R^2, PRESS (वर्गमूल के प्रागुक्त त्रुटिपूर्ण जोड़) का उपयोग करते हुए की गई। कुल गैर वर्गीकरण और पूर्वानुमानों की तुलना RMSE (वर्ग माध्य मूल त्रुटि) और MAPE (माध्य निरपेक्ष प्रतिशत त्रुटि) का उपयोग करते हुए की गई। क्रमसूचक लॉजिस्टिक समाश्रयण आधारित अप्रोच की अपेक्षा बेहतर पाई गई है।

ABSTRACT. The performance of ordinal logistic regression and discriminant function analysis has been compared in crop yield forecasting of wheat crop for Kanpur district of Uttar Pradesh. Crop years were divided into two or three groups based on the detrended yield. Crop yield forecast models have been developed using probabilities obtained through ordinal logistic regression along with year as regressors and validated using subsequent years data. In discriminant function approach two types of models were developed, one using scores and another using posterior probabilities. Performance of the models obtained at different weeks was compared using Adj R^2, PRESS (Predicted error sum of square), number of misclassifications and forecasts were compared using RMSE (Root Mean Square Error) and MAPE (Mean absolute percentage error) of forecast. Ordinal logistic regression based approach was found to be better than discriminant function analysis approach.

Key words – Ordinal logistic regression, Crop yield forecast, Discriminant function analysis.

1. Introduction

Agriculture now-a-days has become highly input and cost intensive. Uncertainties of weather, production, policies, prices, etc. often lead to losses to the farmers. Under the changed scenario today, forecasting of various aspects relating to agriculture is becoming essential. Crop yield forecast is one of the important aspects which needs attention. The techniques employed for forecasting should be able to provide objective, consistent and comprehensible forecasts of crop yield with reasonable precisions well in advance of the harvest. In addition to several agronomic and economic factors, crop yield depends heavily on the vagaries of weather. Therefore, weather based models could be useful in forecasting crop yields.

Various workers have attempted to develop methodology for weather based models for crop yield forecasting such as weather indices based regression models (Agrawal *et al.*, 1986; 2001), discriminant function approach (Agrawal *et al.*, 2012), etc.

Ordinal logistic regression and discriminant function analysis are two approaches which are used for classification of data into various groups and are used for qualitative forecasting. Ordinal logistic regression is a method to describe the effects of some explanatory

variables on a categorical response variable especially when the response variable has an ordinal nature. It is used for prediction of the probability of occurrence of an event by fitting data to a logit function. Discriminant function analysis is another approach which is used for classification of data into various groups and for qualitative forecasting. It is a multivariate technique concerned with separating distinct sets of objects (or sets of observation) & allocating new objects (or observations) to the previously defined groups. The performance of the two methods for classification has been studied by Press and Wilson (1978); O'Gorman and Woolson (1991); Johnson *et al*. (1996); Zibaei and Bakhshoodeh (2008). Ordinal logistic regression has been explored for quantitative forecasting (Kumari and Kumar, 2014).

In this study ordinal logistic regression and discriminant function analysis approaches have been compared for forecasting wheat yield for Kanpur district of Uttar Pradesh both qualitatively and quantitatively.

2. Data

Time series data on yield of wheat crop for Kanpur district of Uttar Pradesh for 39 years (1971-72 to 2009-10) have been obtained from Directorate of Economics and Statistics, Ministry of Agriculture, New Delhi. Weekly weather data for 39 years (1971-72 to 2009-10) of Kanpur district of Uttar Pradesh during the different growth phases of wheat crop have been obtained from Central Research Institute for Dry-Land Agriculture (C.R.I.D.A.), Hyderabad.

3. Methodology

Data from 1971-72 to 2006-07 have been utilized for model development and subsequent three years were used for the validation of the model. Crop years have been divided into two and three groups on the basis of crop yield adjusted for trend effect. The grouping was done into two by taking year having residuals with negative value as bad year (0) and positive values as good year (1) after fitting linear regression between yield and year. Crop years were grouped into three, where residuals (after fitting linear regression between yield and year) have been arranged into ascending order and were divided into three equal groups namely adverse (0), normal (1) and congenial (2). For modeling of crop yield using weather variables in two/three groups, probabilities were obtained by ordinal logistic regression (Agresti, 2002).

Weather variables affect the crop differently during different stages of development, so weekly weather data were used in the study. As weather during pre-sowing period is important for establishment of the crop and also the forecast is required well in advance of harvest, weather data starting from two weeks before sowing, *i.e*., first week of October to about 2 months before harvesting, *i.e*., 15-21 January of the next year has been considered. Thus, the data have been taken up to the first 16 weeks of the crop cultivation which included 40th standard meteorological week (SMW) to 52nd SMW of a year and 1st SMW to 3rd SMW of the next year. Using 5 variables in 16 weeks as such makes number of explanatory variable 80 in ordinal logistic regression giving rise to the problem of number of explanatory variables (80) more than the number of observations. The following strategy has been used to solve the problem. At first week, the weather variables corresponding to the pre-defined groups have been used to compute probabilities by stepwise ordinal logistic regression. At the second week, the weather variables of this week along with probabilities computed at the first week have been used to compute probabilities in second week using stepwise ordinal logistic regression. These steps have been repeated in third week as well and so on upto last week. Forecast models were fitted at different weeks starting from 52nd SMW using stepwise regression procedure, taking probabilities at week of forecast along with year as regressors. Besides this discriminant function analysis has been carried out using weather variables in the pre-defined groups (Agrawal *et al.*, 2012) in similar procedure as followed for logistic. Two types of forecast models using discriminant function approach were developed, one using scores and the other one using probabilities as regressors. The comparison of these two approaches has been performed.

3.1. *Modeling with two and three groups*

Probability of good year, *i.e*., $P_1 = P(Y = 1)$ under ordinal logistic regression with multiple explanatory variables, $x = (x_1, \ldots, x_p)$ of p predictors is given by:

$$P_i = \frac{\exp\left(\alpha + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_p x_p\right)}{1 + \exp\left(\alpha + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_p x_p\right)}$$

where, α is the intercept and β's are the regression coefficients. Thus, probability of bad year is $P\,(Y = 0) = 1\text{-}P_1$. Forecast model was obtained taking P_1 and year (T) as regressors.

When dependent variable has an ordinal nature, *i.e*., taking three values zero, one, two then the ordered multiple response models assume the relationship:

$$\text{logit}\left[P\left(Y \le j-1 \middle| x\right)\right] = \alpha_j + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_p x_p,\ j = 1,2$$

TABLE 1

Wheat yield forecast models for different week alongwith comparison of forecast for two groups

Week of forecast (SMW)	Forecast Models				Comparison of forecasts	
	Forecast regression equation	Adj R^2	PRESS	Mis.	RMSE	MAPE
52	Yield = -1120.18 + 3.93 P_1 + 0.57 T (66.50) (0.74) (0.03)	0.903**	168.59	2	2.25	6.11
1	Yield = -1118.89 + 3.88 P_1 + 0.57 T (68.350) (0.78) (0.034)	0.898**	178.92	2	2.31	6.34
2	Yield = -1118.66 + 3.85 P_1 + 0.57 T (69.47) (0.80) (0.03)	0.894**	185.77	2	2.36	6.51
3	Yield = -1118.71 + 3.85 P_1 + 0.57T (70.00) (0.82) (0.03)	0.892**	189.26	2	2.38	6.61

Note : Figures in brackets denote Standard Error of regression coefficients, ** Significant at p = 0.01, Mis.= Misclassifications

Thus, ordinal logistic regression model is given as:

$$P_0 = \exp\left(\alpha_1 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_p x_p\right) \Big/ \left[1 + \exp\left(\alpha_1 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_p x_p\right)\right]$$

$$P_0 + P_1 = \exp\left(\alpha_2 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_p x_p\right) \Big/ \left[1 + \exp\left(\alpha_2 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_p x_p\right)\right]$$

and $P_0 + P_1 + P_2 = 1$

where, P_0 is probability of $Y = 0$, P_1 is probability of $Y = 1$ and P_2 is probability of $Y = 2$, α_j's are the intercepts and β_i's are the regression coefficients. Forecast model was obtained taking P_1, P_2 and year (T) as regressors.

For validation of models, forecasts of subsequent years were obtained and Root Mean Square Error (RMSE) & Mean Absolute Percentage Error (MAPE) of the forecasts were computed. Different models have been compared using Adjusted R^2 (Adj R^2), Predicted error sum of square (PRESS), RMSE and MAPE of forecast.

4. Results and discussion

Yield data (1971-72 to 2006-07) have been classified into two groups and three groups using detrended yield. Out of thirty six years, the numbers of good (1) crop years were twenty and bad (0) were sixteen and the numbers of years congenial (2), normal (1) or adverse (0) were twelve.

Weather variables in two groups were used to obtain probability of good year through stepwise logistic regression. Regression models have been fitted using stepwise regression by taking yield as the dependent variable and the probability of good year ($Y = 1$) and year as the regressors for different weeks of forecast starting from 52nd SMW. Wheat yield forecast models alongwith Adj R^2, PRESS and number of misclassifications for different weeks are given in Table 1.

Results show that Adj R^2 varied from 0.898 to 0.903. It explained about 90.30% variability in the yield. PRESS varied from 168.59 in 52nd SMW to 189.26 in 3rd SMW. So, PRESS value was minimum for 52nd week. Therefore, 52nd week has best model fit as compared to others. Further number of misclassifications were two (1991-92 and 1995-96) in all the weeks. Evaluation of the forecasts of subsequent years at different weeks has been done by RMSE, MAPE and number of misclassifications of forecasts. RMSE varied from 2.25 in 52nd SMW to 2.38 in 3rd SMW. Also, MAPE ranged from 6.11 in 52nd SMW to 6.61 in 3rd SMW. So, RMSE and MAPE values were minimum for 52nd week. There was no misclassification in the forecast years at different weeks.

TABLE 2

Wheat yield forecast models for different week alongwith comparison of forecast for three groups

Week of forecast (SMW)	Forecast models				Comparison of forecasts	
	Forecast regression equation	Adj R^2	PRESS	Mis.	RMSE	MAPE
52	Yield = -1146.70 + 4.77 P_1 + 6.45 P_2 + 0.57 T (61.41) (2.20) (1.05) (0.03)	0.918**	147.45	9	3.05	8.11
1	Yield = -1179.13 + 4.46P_1 + 6.39P_2 + 0.60 T (60.68) (1.89) (1.05) (0.03)	0.923**	138.02	10	3.33	9.69
2	Yield = -1179.55 + 4.24 P_1 + 6.24 P_2 + 0.60 T (60.68) (1.82) (0.97) (0.03)	0.924**	136.36	8	3.42	9.80
3	Yield = -1177.54 + 4.20 P_1 + 6.19 P_2 + 0.60 T (60.43) (1.80) (0.94) (0.03)	0.924**	135.06	8	3.48	9.84

Note : Figures in brackets denote Standard Errors,** Significant at p = 0.01, Mis. = Misclassifications

TABLE 3

Comparison between ordinal logistic regression and discriminant function using RMSE and MAPE

Week of forecast (SMW)	Two Groups						Three Groups					
	Logistic		Discriminant (scores)		Discriminant (probability)		Logistic		Discriminant (scores)		Discriminant (probability)	
	RMSE	MAPE	RMSE	MAPE	RMSE	MAPE	RMSE	MAPE	RMSE	MAPE	RMSE	MAPE
52	2.25	6.11	5.00	13.80	4.15	10.69	3.05	8.11	4.38	13.06	4.44	10.50
1	2.31	6.34	4.81	13.57	4.29	11.09	3.33	9.69	4.41	13.23	4.46	10.57
2	2.36	6.51	4.74	13.37	4.24	10.99	3.42	9.80	4.41	13.23	4.50	10.87
3	2.38	6.61	4.74	13.37	4.21	10.95	3.48	9.84	4.41	13.23	4.53	11.06

Using weather variables in three groups probabilities have been obtained by stepwise logistic regression. Regression models have been fitted using stepwise regression by taking yield as the dependent variable and the probabilities (normal and congenial years) and year as the regressors. Wheat yield forecast models for different weeks alongwith Adj R^2, PRESS and number of misclassifications alongwith comparison of forecast using RMSE and MAPE are given in Table 2.

Results show that Adj R^2 varied from a minimum 0.918 in 52nd SMW to a maximum of 0.924 in 3rd SMW. PRESS varied from 135.06 in 3rd SMW to 147.45 in 52nd SMW. It was observed that number of years misclassified varied from 8 at 2nd and 3rd SMW to 10 at 1st SMW. Thus, 3rd week gave minimum PRESS and number of misclassifications alongwith maximum Adj R^2. Therefore 3rd week has best model fit as compared to others. It explained about 92.40% variability in the model.

TABLE 4

Observed and forecasts of subsequent years at 52nd SMW

Week of forecast (SMW)	Year	Observed yield	Forecasted yield	% Deviation of forecast
52	2007-08	30.08	33.31	10.74
	2008-09	33.56	33.88	0.95
	2009-10	32.31	34.45	6.62

Evaluation of the forecasts developed at different weeks has been done by RMSE, MAPE and misclassifications. RMSE varied from 3.05 in 52nd SMW to 3.48 in 3rd SMW. Also, MAPE ranged from 8.11 in 52nd SMW to 9.84 in 3rd SMW. So, RMSE and MAPE values were minimum for 52nd week. Number of misclassification in the forecast years was one for all weeks.

Ordinal logistic regression and discriminant function have been compared qualitatively on the basis of number of misclassifications and quantitatively on the basis of RMSE and MAPE of forecast. Results of comparison for logistic regression approach and discriminant function approach are given in Table 3.

Table 3 revealed that RMSE under logistic regression approach were less as compared to discriminant function analysis in both two and three groups cases. Further RMSE was minimum for 52nd SMW under logistic regression approach in two groups case. Also, MAPE under logistic regression approach was less as compared to discriminant function analysis in both two and three groups cases. Further MAPE was minimum for 52nd SMW under logistic regression approach in two groups case.

The number of misclassifications in forecast years was two for discriminant function analysis using scores in three group case. There was one misclassification for discriminant function analysis using scores in two group case and one misclassification for discriminant function analysis using probability in two and three group cases. Further under two groups logistic regression there was no misclassification in forecast years for all weeks of forecast.

Observed and forecasts of subsequent years using ordinal logistic regression model for 52nd SMW is given in Table 4.

5. Conclusion

The performance of ordinal logistic regression approach is better than discriminant function analysis for forecasting crop yield. Two group classification is better than three groups. Two groups classification with ordinal logistic regression gave best forecasts. Reliable forecast of crop yield can be obtained using ordinal logistic regression model along with trend at 52nd SMW, *i.e.*, 13th period which is 11 weeks after sowing. Thus, model based on ordinal logistic regression with two groups classification at 52nd week can be recommended for forecasting crop yield.

References

Agrawal, R., Jain, R. C. and Jha, M. P., 1986, "Models for studying rice crop-weather relationship", *Mausam*, **37**, 1, 67-70.

Agrawal, R., Jain, R. C. and Mehta, S. C., 2001, "Yield forecast based on weather variables and agricultural inputs on agroclimatic zone basis", *Indian Journal of Agricultural Science*, **71,** 7**,** 487-490.

Agrawal, R., Chandrahas and Aditya, K., 2012, "Use of discriminant function anaylsis for forecasting crop yield", *Mausam*, **63**, 3, 455-458.

Agresti, A., 2002, "Categorical data analysis", New York : Wiley.

Johnson, D. A., Alldredge, J. R. and Vakoch, D. L., 1996, "Potato late blight forecasting models for the semiarid-environment of south-central Washington", *American phytopathology,* **86**, 480-84.

Kumari, V. and Kumar A., 2014, "Forecasting of wheat (Triticumaestivum) yield using ordinal logistic regression", *Indian Journal of Agricultural Science*, **84**, 6, 691-694.

O'Gorman, T. W. and Woolson, R. F., 1991, "Variable selection to discriminant between two groups : stepwise logistic regression or stepwise discriminant analysis?", *American Statistical Association*, **45**, 3, 187-93.

Press, S. J. and Wilson, S., 1978, "Choosing between Logistic Regression and Discriminant Analysis", *Journal of the American Statistical Association*, **73**, 364, 699-705.

Zibaei, M. and Bakhshoodeh, M., 2008, "Investigating Determinants of Sprinkler Irrigation Technology Discontinuance in Iran : Comparison of Logistic Regression and Discriminant Analysis", *Am- Euras. J. Agric. & Environ. Sci.*, **2**, 46-50.