Indian Journal of Agricultural Sciences **84** (6): 691–4, June 2014/Article

Forecasting of wheat (*Triticum aestivum*) yield using ordinal logistic regression

VANDITA KUMARI[1] and AMRENDER KUMAR[2]

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

Received: 9 September 2013; Revised accepted: 11 February 2014

ABSTRACT

In this study, uses of ordinal logistic model based on weather data has been attempted for forecasting wheat (*Triticum aestivum* L.) yield in Kanpur district of Uttar Pradesh. Weekly weather data (1971-72 to 2009-10) on maximum temperature, minimum temperature, morning relative humidity, evening relative humidity and rainfall for 16 weeks of the crop cultivation along with the yield data of wheat crop have been considered in the study. Crop years were divided into two and three groups based on the detrended yield. Yield forecast models have been developed using probabilities obtained through ordinal logistic regression along with year as regressors for different weeks. Data from 1971-72 to 2006-07 have been utilized for model fitting and subsequent three years (2007-08 to 2009-10) were used for the validation of the model. Evaluation of the performance of the models developed at different weeks has been done by Adj R^2, PRESS (Predicted error sums of squares) and number of misclassifications. Evaluation of the forecasts were done by RMSE (Root mean square error) and MAPE (Mean absolute percentage error) of forecast.

Key words: Ordinal logistic regression, Wheat yield forecast

Timely and effective pre-harvest forecast of the crop yield is important for advance planning, formulation and implementation of policies related to the crop procurement, distribution, price structure and import export decisions etc. These are also useful to farmers to decide in advance their future prospects and course of action. The yield of any crop is affected by technological change and weather variability. It can be assumed that the technological factors will increase yield smoothly through time and, therefore, year or some other parameters of time can be used to study the overall effect of technology on crop yield. Weather variability both within and between seasons is the second and uncontrollable source of variability in yields. Therefore, model based on weather and year as explanatory variables can be used for forecasting crop yield. Weather variables affect the crop differently during different stages of development. Thus extent of weather influence on crop yield depends not only on the magnitude of weather variables but also on the distribution pattern of weather over the crop season which, as such, calls for the necessity of dividing the whole crop season into finer intervals and studying crop weather relationships in these intervals. However, doing so will increase number of variables in the model and in turn a large number of parameters will have to be evaluated from the data and sufficient number of observations may not be available for precise estimation of these parameters. Thus, a technique based on relatively smaller number of manageable parameters and at the same time taking care of entire weather distribution may solve the problem.

Various workers have attempted to develop methodology for weather based models of crop yields forecasting such as weather indices based regression models (Agrawal *et al.* 1980, 1983, 2001; Jain *et al.* 1980, Chandrahas *et al*. 2010), discriminant function approach (Rai and Chandrahas 2000, Chandrahas *et al.* 2010, Agrawal *et al.* 2012), water balance technique (Saksena *et al*. 2001), complex polynomials using GMDH technique (Mehta *et al.* 2010).

In present investigation use of ordinal logistic regression has been explored for crop yield forecasting.

MATERIALS AND METHODS

Time series data on yield of wheat (*Triticum aestivum* L.) 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.

Wheat is generally sown in the end of October when average daily temperature is around 23-25°C. Germination takes 6-7 days after sowing of the crop or near about one week. Then crown root initiation occurs at 20-25 days after sowing or in about 3 weeks from germination. Tillering phase starts just after the crown root initiation phase and lasts up to 40-45 days after sowing or nearly about 2-3 weeks after crown root initiation phase. Jointing is the peak plant growth stage and starts after the tillering phase or 45-60 days after sowing. In this phase the vegetative growth of

[1] Ph D Scholar (e mail: vandita.iasri@gmail.com), Division of Sample Survey; [2] Senior Scientist (e mail: akjha@iari.res.in), Agricultural Knowledge Management Unit (AKMU), Indian Agricultural Research Institute, Pusa, New Delhi 110 012

the crop is completed and the crop then goes to the reproductive phase. The reproductive phase lasts 60-85 days after sowing or near about 4-5 weeks after the jointing phase.

Weekly weather data (1971-72 to 2009-10) on weather variables, viz. maximum temperature, minimum temperature, morning relative humidity, evening relative humidity and rainfall for Kanpur district of Uttar Pradesh, have been obtained from Central Research Institute for Dryland Agriculture (CRIDA), Hyderabad. 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 on five weather variables for 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 next year has been used in the study.

Crop years have been divided into two groups namely good (1) and bad (0) and three groups namely adverse (0), normal (1) and congenial (2) on the basis of crop yield adjusted for year effect. The grouping was done into two by taking year having residuals with negative value as zero and positive values as one 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 divided into three equal groups namely adverse (0), normal (1) and congenial (2). The data from 1971-72 to 2006-07 have been utilized for model fitting and remaining three years (2007-08 to 2009-10) were utilized for the validation of the models. Using weather variables in these two and three groups, probabilities were obtained by ordinal logistic regression. These probabilities along with year as regressors were used for development of forecast model. The models were developed using stepwise regression procedure.

Use of five variables in 16 weeks as such makes number of explanatory variable 80 in ordinal logistic regression. Thus, the following strategy has been used to solve the problem of number of variables more than number of data points. At first week (40th SMW), the weather variables corresponding to the pre-defined groups have been used to compute probabilities by stepwise logistic regression. At the second week (41st SMW), the weather variables of this week along with probabilities computed at the first week have been used to compute probabilities using stepwise logistic regression. These steps have been repeated in third week as well and so on up to last week (3rd SMW). Forecast models were developed at different weeks starting from 52nd SMW (through stepwise regression procedure) taking probabilities obtained at corresponding week of forecast along with year as regressors.

Modelling with two groups

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

$$P_1 = \frac{\exp(\alpha+\beta_1 x_1+\beta_2 x_2+...+\beta_p x_p)}{1+\exp(\alpha+\beta_1 x_1+\beta_2 x_2+...+\beta_p x_p)}$$

where α is the intercept and β's are the regression coefficients.

Thus, $P(Y=0) = 1- P_1$

Forecast model

Yield forecast models were fitted using stepwise regression at different weeks starting from 52nd SMW taking probability P_1 at week of forecast (obtained through stepwise logistic regression model) along with year as regressors. The model fitted was

$$\text{Yield} = a + b_1 P_1 + b_2 T + \varepsilon$$

where, a is intercept of the model, b_i's are the regression coefficients, P_1 is the probability good year (Y=1), T is year, ε is error ~ N $(0,\sigma^2)$.

Modelling with three groups

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

$$\text{logit}\,[P(Y \leq j-1 | \mathbf{x}) = \chi_j + \beta_1 x_1 + \beta_2 x_2 + \beta_p x_p,\ j=1, 2$$

The ordinal logistic regression model is given as:

$$P_0 = \frac{\exp(\alpha_1 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_p x_p)}{1+\exp(\alpha_1 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_p x_p)}$$

$$P_0 + P_1 = \frac{\exp(\alpha_2 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_p x_p)}{1+\exp(\alpha_2 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_p x_p)}$$

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.

Yield forecast models were fitted using stepwise regression at different weeks starting from 52nd SMW taking probabilities P_1 and P_2 at week of forecast along with year as regressors. The model fitted was

$$\text{Yield} = a + b_1 P_1 + b_2 P_2 + b_3 T + \varepsilon$$

where, P_1 and P_2 are the probabilities of Y=1 and Y=2. Other symbols are same as defined earlier.

For validation of models, forecasts of subsequent years were obtained and RMSE and MAPE of the forecasts were computed.

Different models have been compared using following procedures. (i) Adjusted R^2

Adjusted R^2 is given by the following formula:

$$R_{adj}^{\ 2} = 1 - \frac{ss_{res}/(n-p)}{ss_t/(n-1)}$$

where, $ss_{res}/(n-p)$ is the residual mean sum of square and $ss_t/(n-1)$is the total mean sum of square.

(ii) Predicted error sum of square (PRESS)

The (PRESS) statistic is defined as:

$$PRESS = \sum_{m=1}^{n} \left[Y_m - \hat{Y}_m^{(m)} \right]^2$$

where, Y_m is the value of dependent variable of m^{th} observation (crop yield in the present study of m^{th} year) and $\hat{Y}_m^{(m)}$ is the forecast of Y_m computed from a model fitted without the m^{th} data point. It is generally regarded as how well a regression model will perform in predicting new data.

(iii) Root Mean Square Error (RMSE)

The formula of (RMSE) of forecast is given as:

$$RMSE = \left[\left\{ \frac{1}{n} \sum_{m=1}^{n} (O_m - F_m)^2 \right\} \right]^{\frac{1}{2}}$$

where, O_m and F_m are the observed and forecasted values of the crop yield respectively and n is the number of years for which forecasting has been done.

(iv) Mean absolute percentage error (MAPE)

The formula of (MAPE) of forecast is given as:

$$MAPE = \frac{100}{n} \sum_{m=1}^{n} \left| \frac{F_m - O_m}{O_m} \right|$$

where, O_m and F_m are the observed and forecasted value of the crop yield respectively and n is the number of years for which forecasting has been done.

RESULTS AND DISCUSSION

Linear regression was fitted between yield and year. The fitted equation was

Yield= –1127.188 + 0.579 T

Observed and predicted values of yield are presented in Fig 1.

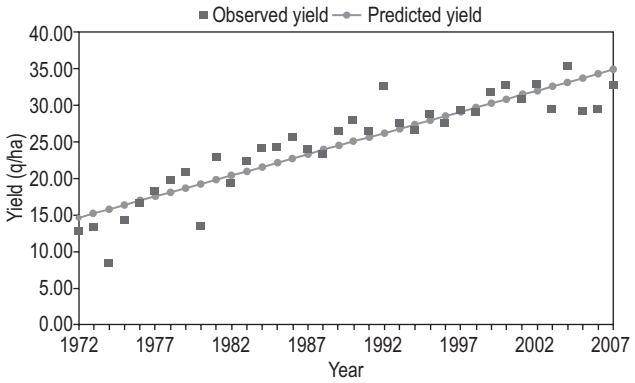

Fig 1 Observed and predicted wheat yield (1971-72 to 2006-07)

Forecast model - Two groups case

Yield data (1971-72 to 2006-07) have been classified into two groups namely good (1) and bad (0) on the basis of crop yield adjusted for the year effect taking negative residuals as bad years and positive residuals as good years. Out of thirty six years, the numbers of good (1) crop years are twenty and bad (0) are sixteen. 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 is given in Table 1.

Table 1 Wheat yield forecast models for different weeks

Week of forecast (SMW)	Forecast regression equation	Adj R^2	PRESS	Misclassifications
52	Yield = -1120.18 + 3.93 P + 0.57 T	0.9030**	168.59	2
1	Yield = -1118.89 + 3.88 P + 0.57 T	0.8976**	178.91	2
2	Yield = -1118.66 + 3.85 P + 0.57 T	0.8942**	185.76	2
3	Yield = -1118.71 + 3.85 P + 0.57T	0.8926**	189.26	2

**Significant at p = 0.01

This table revealed that that the Adj R^2 varied from a minimum 0.8926 in 3rd SMW to a maximum of 0.9030 in 52nd SMW. The model explained about 90.30% variability in the yield. PRESS varied from 168.595 in 52nd SMW to 189.263 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.

Observed and forecasts of subsequent years for 52nd week using these model are given in Table 2.

Table 2 Observed and forecast for different years at 52nd SMW

Week of forecast (SMW)	Year	Observed yield	Forecasted yield	% deviation of forecast	95% confidence interval	
					Lower limit	Upper limit
52	2007-08	30.08	33.31	10.74	28.75	37.86
	2008-09	33.56	33.88	0.95	29.30	38.45
	2009-10	32.31	34.45	6.62	29.86	39.05

Evaluation of the forecasts of subsequent years at different weeks has been done by RMSE, MAPE and number of misclassifications of forecasts which are given in Table 3.

Table 3 Comparison of forecasts in different weeks

Week of forecast (SMW)	RMSE	MAPE	Misclassification
52	2.25	6.11	0
1	2.31	6.34	0
2	2.36	6.51	0
3	2.38	6.61	0

The Table 3 indicates that 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.

Forecast model - Three groups case

Yield data (1971-72 to 2006-07) have been classified into three groups namely congenial, normal or adverse on the basis of crop yield adjusted for the trend each group consisting of twelve years. Using weather variables in these 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 and year as the regressors. Wheat yield forecast models for different weeks are given in Table 4.

Table 4 Wheat yield forecast models for different weeks

Week of forecast (SMW)	Forecast regression equation	Adj R^2	PRESS	Mis-classifi-cations
52	Yield= -1146.70 + 4.77 P_1 + 6.45 P_2 + 0.57 T	0.9181**	147.45	9
1	Yield= -1179.13 + 4.46 P_1 + 6.39P_2 + 0.60 T	0.9227**	138.02	10
2	Yield= -1179.55 + 4.24 P_1 + 6.24 P_2 + 0.60 T	0.9236**	136.36	8
3	Yield= -1177.54 + 4.20 P_1 + 6.19 P_2 + 0.60 T	0.9242**	135.06	8

** Significant at p = 0.01

The Table 4 revealed that Adj R^2 varied from a minimum 0.9181 in 52nd SMW to a maximum of 0.9242 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, number of misclassifications and maximum Adj R^2.

Observed and forecasts of subsequent years for 52nd week using these model are given in Table 5.

Table 5 Observed and forecast for different years at 52nd SMW

Week of forecast (SMW)	Year	Obser-ved yield	Fore-casted yield	% deviation of forecast	95% confidence interval	
					Lower limit	Upper limit
52	2007-08	30.08	33.22	10.47	29.03	37.42
	2008-09	33.56	33.81	0.75	29.60	38.02
	2009-10	32.31	36.54	13.09	32.22	40.86

Table 6 Comparison of forecast at different weeks

Week of forecast (SMW)	RMSE	MAPE	Misclassification
52	3.05	8.11	1
1	3.33	9.69	1
2	3.42	9.80	1
3	3.48	9.84	1

Evaluation of the forecasts developed at different weeks has been done by RMSE, MAPE and misclassifications which are given in Table 6.

The Table 6 revealed that, 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 was same for all weeks.

Comparison between two and three groups

Comparison between two and three groups indicated three groups case gave better fit to model as compared to two groups case in terms of Adj R^2 and PRESS but forecasts of two groups case were better.

CONCLUSION

On the basis of above results, it can be concluded that reliable forecast of crop yield can be obtained using model taking probabilities through ordinal logistic regression along with year as explanatory variables. Appropriate time of forecast is 52nd SMW, i e 11 weeks after sowing.

REFERENCES

Agrawal R, Jain R C and Singh D. 1980. Forecasting of rice yield using climatic variables. *Indian Journal of Agricultural Sciences* **50**(9): 680–4.

Agrawal R, Jain R C and Jha M P. 1983. Joint effects of weather variables on rice yields. *Mausam* **34**(2): 177–81.

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).

Agrawal R, Chandrahas and Aditya K. 2012. Use of discriminant function anaylsis for forecasting crop yield. *Mausam* **63**(3): 455–8.

Chandrahas, Agrawal R and Walia S S. 2010. Use of discriminant function and principal component techniques for weather based crop yield forecasts. (IASRI publication).

Jain R C, Agrawal R and Jha M P. 1980. Effect of climatic variables on rice yield and its forecast. *Mausam* **31**(4): 591–6.

Mehta S C, Pal S and Kumar V. 2010. Weather based model for forecasting potato yield in Uttar Pradesh. (IASRI publication).

Rai T and Chandrahas. 2000. Use of discriminant function of weather parameters for developing forecast model of rice crop, IASRI publication.

Saksena Asha, Jain R C and Yadav R L. 2001. Development of early warning and yield assessment models for rainfed crops based on agro-meteorological indices. (IASRI Publication).