*Journal of Agrometeorology 21 (2) : 188-192 (June 2019)*

# Comparison of Lasso and stepwise regression technique for wheat yield prediction

**SUDHEER KUMAR, S.D. ATTRI and K.K. SINGH**

*Agromet Advisory Service Division*
*India Meteorological Department-New Delhi*
*Email-sudheergkp@gmail.com*

## ABSTRACT

Multiple regression approach has been used to forecast the crop production widely. This study has been undertaken to evaluate the performance of stepwise and Lasso (Least absolute shrinkage and selection operator) regression technique in variable selection and development of wheat forecast model for crop yield using weather data and wheat yield for the period of 1984-2015, collected from IARI, New Delhi. Statistical parameters viz. $R^2$, RMSE, and MAPE were 0.81, 195.90 and 4.54 per cent respectively with stepwise regression and 0.95, 99.27, 2.7 percentage, respectively with Lasso regression. Forecast models were validated during 2013-14 and 2014-15. Prediction errors were -8.5 and 10.14 per cent with stepwise and 1.89 and 1.64 percent with the Lasso. This shows that performance of Lasso regression is better than stepwise regression to some extent.

***Keywords:*** Weather variables, stepwise regression, lasso regression, cross validation, forecast model

Reliable and routine forecast of crop production play a vital role for advance planning, formulation and implementation of a number of policies dealing with food procurement its distribution, pricing structure, import and export decisions and for exercising several administrative measures related to the storage and marketing of the agricultural commodities. Many techniques have been developed to forecast crop production. However, two main approaches to predicting crop yield based on weather conditions are simulation models and multiple regression models (Lee, 2011). The simulation models designed to forecast crop yield use detailed crop biology and requires extensive information on soil type, plant parameters and weather data related to the crop development stage, which is often not readily available (Walker 1989). Tannura *et al.* (2008) and other studies have proven that multiple regression models have high explanatory power and can represent relationships between weather conditions and crop yield. Thus, the multiple regression models based on agrometeorological parameters are used to predict the crop production which is not only easier to use, it is also likely to be more accurate than the simulation model approach. Agrawal *et al.* (2012) have developed forecast models for wheat yield in Kanpur district using discriminant function analysis of weakly data on weather variables. Rai and Chandrahas (2000) used discriminant function analysis of weather variables to develop a statistical model for the pre-harvest forecast of rice yield in Raipur district of Chhattisgarh.

Variable selection in multi regression model plays important role in prediction accuracy as (i) it makes the model easier to interpret, removing variables that are redundant and do not add any information (ii) reduces the size of the problem to enable algorithms to work faster, making it possible to handle with high-dimensional data and (iii) reduces over fitting in model.Many techniques have been developed for selection of predictor variables such as ordinary least square (OLS), stepwise regression, ridge regression, Lasso regression and elastic net regression.

Stepwise variable selection methods (Kutner *et al.*, 2004) based on significance tests are widely used in medical and research field. One of the main issues with stepwise regression is that it searches a large space of possible models. Hence it is prone to overfitting the data. Other issue is that when estimating the degrees of freedom, the number of the candidate independent variables from the best fit selected may be smaller than the total number of final model variables, causing the fit to appear better than it is when adjusting the $r^2$ value for the number of degrees of freedom.

Another variable selection Lasso (Least Absolute Shrinkage and Selection Operator) - was first formulated by Tibshirani(1996). It is a powerful method that performs two main tasks viz. regularization and feature selection in order to enhance the prediction accuracy and interpretability of the statistical model. Regularization techniques are used to prevent statistical overfitting in a predictive model.In general statistics number of observation should be greater than number of variable, where the number of variables is bigger than the number of observations, the selection of variable gains greater importance. In this situation, Lasso regression can perform with some limitation as compared to stepwise regression. Therefore, the aim of this study is to evaluate the best performance of stepwise and Lasso variable selection methods in order to enhance accuracy of crop yield forecast.

## MATERIALS AND METHODS

The daily weather data such as maximum and minimum temperature, morning and evening relative humidity, rainfall and wheat yield from 1984 to 2015 were collected from IARI, New Delhi.

### *Yield forecast model*

The wheat crop yield forecast model was developed using the statistical model following Ghosh *et al.* (2014).

$$Y = A_0 + \sum_{i=1}^{p} \sum_{j=0}^{1} a_{ij} Z_{ij} + \sum_{i \neq i=1}^{p} \sum_{j=0}^{1} a_{iij} Z_{iij} + cT + e \qquad (1)$$

where,

$$Z_{ij} = \sum_{w=1}^{m} r_{iw}^{j} X_{iw}\ and Z_{iiw} = \sum_{w=1}^{m} r_{iiw}^{j} X_{iw} X_{iw}$$

Here Y is the wheat yield (kg ha$^{-1}$)

'$r_{iw}$' is correlation coefficient of yield with i-th weather variable in w-th period

'rii'w' is correlation coefficient (adjusted for trend effect) of yield with product of i-th and i'-th weather variables in w-th period

'p' is number of weather variables used

'm' is period of forecast

'e' is error term

For each weather variable, two variables were generated- one as a weighted accumulation of weekly data on weather variable, weights being the correlation coefficients of the weather variables, in respective weekly with yield and another one as unweighted, it is the simple accumulation of weather variable. Similarly, for the joint effect of weather variables, weekly interaction variables were generated using weekly products of weather variables taking 2 at a time. Weather indices were denoted as Z with 0 as unweighted indices and 1 as weighted indices. As we are taking Tmax as a 1st variable, hence weather indices of Tmax for unweighted is Z10 and Tmax weighted is Z11. In this way, thirty indices have been prepared from individual and joint variable for all five weather variables.

### *Stepwise and Lasso regression*

Stepwise variable and Lasso variable selection method were taken in to consideration for comparison of their performances for wheat crop yield. Stepwise regression is a combination of the forward and backward selection techniques and variable selection is carried out by automatic method. All predictor variables in the model are checked to see if their significance has been reduced below the specified tolerance level, it added in final model. If a variable is found nonsignificant, it is removed from the model.

Lasso is a powerful method that performs two main tasks: regularization and variable selection. The Lassosets a constraint on the sum of the absolute values of the model parameters; the sum has to be less than a fixed value (upper bound). In order to do so the method apply a shrinking process where it penalizes the coefficients of the regression variables shrinking some of them to zero. During variable selection process the variables that still have a non-zero coefficient after the shrinking process are selected to be part of the model. The purpose of this process is to minimize the prediction error.

The tuning parameter λ plays important role that controls the strength of penalty in variable selection. The larger is the parameter λ the more number of coefficients are shrunk to zero, in this way variables can be reduced. The smaller λ value trends to more coefficient, in case if $\lambda=0$ it becomes Ordinary Least Square Error regression. In terms of the tuning parameter λ when λ increases, bias increases and variance decreases, indeed a trade-off between bias and variance has to be found. Moreover the Lasso helps to increase the model interpretability by eliminating irrelevant variables that are not associated with the response variable; this way also over fitting is reduced.

The Lasso technique solves this regularization problem. For a given value of λ, a nonnegative parameter, Lasso solves the problem byTibshirani (1997):

$$\min_{\beta_0 \beta} \left( \frac{1}{2N} \sum_{i=1}^{N} (Y_i - \beta_0 - X_i^T \beta)^2 + \lambda \sum_{j=1}^{p} |\beta_j| \right) \qquad (2)$$

Where

N is the number of observations;

$y_i$ is the response at observation $i$.

**Table1:** Stepwise regression coefficients for variable selection

| Model | Unstandardized coefficients | | Standardized coefficients | t | Sig. |
|---|---|---|---|---|---|
| | B | Std. error | Beta | | |
| (Constant) | 1492.35 | 2.63 | | 19.43 | 0.0000 |
| Time | 51.17 | 0.11 | 1.02 | 3.85 | 0.0009 |
| Z141 | 0.42 | 0.034 | 0.22 | 4.80 | 0.0001 |
| Z351 | 0.16 | 309.83 | 0.24 | 4.81 | 0.0001 |

**Table 2:**Lasso regression coefficients for variable selection

| Variable | Coefficient |
|---|---|
| Constant | 1309.26 |
| Time | 50.025 |
| Z20 | -1.51 |
| Z41 | 11.99 |
| Z141 | 0.38 |
| Z350 | 0.011 |
| Z351 | 0.0788 |

**Table 3:** Comparative results of stepwise and Lasso regression

| Parameter | Stepwise regression | Lasso regression |
|---|---|---|
| R$^2$ | 0.81 | 0.95 |
| RMSE | 195.90 | 99.27 |
| MAPE | 4.54 | 2.30 |

$x_i$ is data, a vector of $p$ values at observation $i$.

$\lambda$ is a nonnegative regularization parameter corresponding to one value of Lambda.

The parameters $\beta_0$ and $\beta$ are scalar and $p$-vector respectively.

As $\lambda$ increases, the number of non zero components of $\beta$ decreases

The Lasso problem involves the $L^1$ norm of $\beta$, as contrasted with the elastic net algorithm.

## RESULTS AND DISCUSSION

### *Stepwise regression*

Stepwise regression was performed using SPSS statistical software for the selection of significant variable and results of the coefficient are presented in Table 1. From the table, Z141, Z351 are the significant variables. The yield forecast equation has been developed using the significant generated weather variables. The final yield forecast model using weather variables has been presented in equation 3.

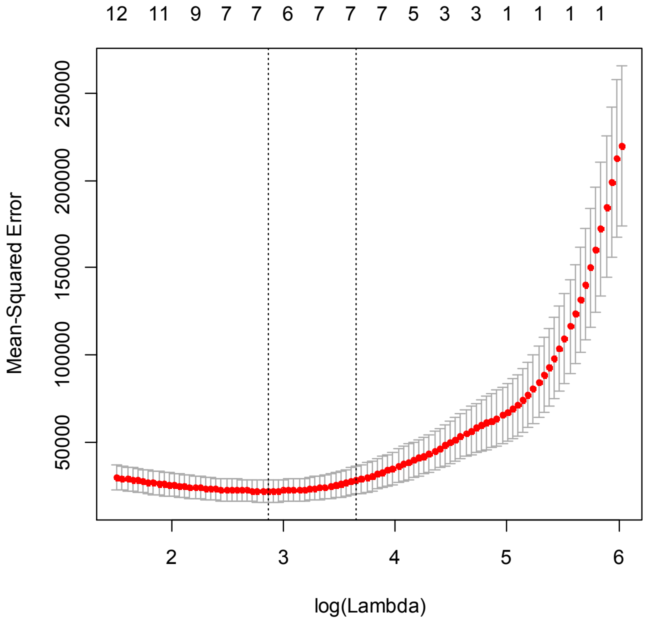

**Fig.1:**Cross validation for minimum mean square error

Yield=1492.35+0.423*Z141+0.163*Z351+51.17*Time (3)

### *Lasso regression*

Fig. 1 shows the threshold value of lambda with upper and lower limit. Two selected $\lambda$ are indicated by the vertical dotted lines. Lambda value can be selected between two vertical dotted lines. Lambda value used to generate the weather significant variables and it was shown in Table 2. Z20, Z41, Z141, Z350 and Z351 variables are found most significant and these variables were used to develop the forecast model. Forecast model using lasso variable selection method was shown in equation 4.

Yield=1309.26-1.51*Z20+11.98*Z41+0.38*Z141
+0.011*Z350+0.038*Z351+50.025*Time (4)

### *Comparison of stepwise and Lasso regression*

The comparative values of statistical parameters, R$^2$, RMSE and MAPE are shown in the Table 3. It is clears that, all statistical parameters of Lasso variable selection method was better than the stepwise method. Fig.2 shows the

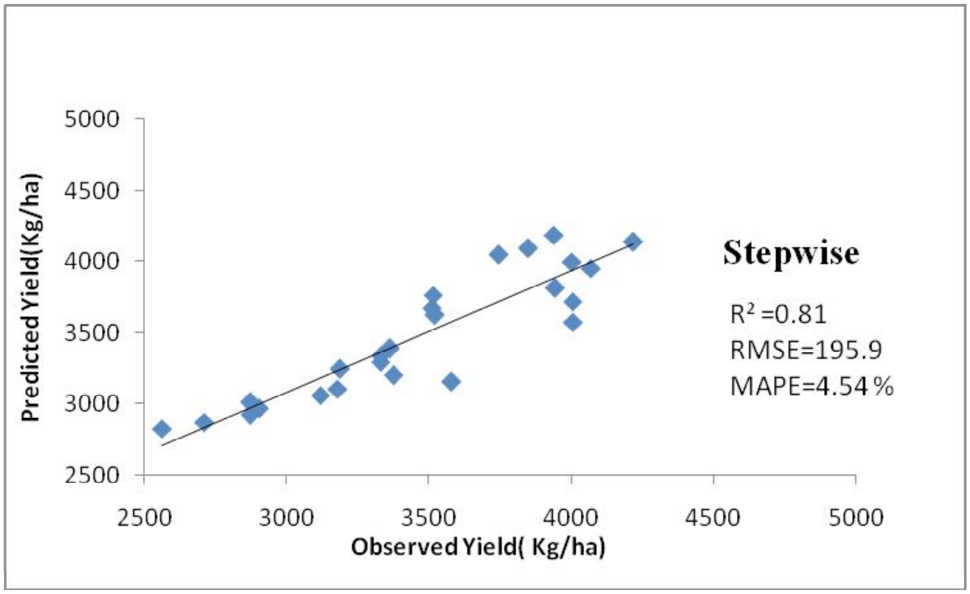

**Fig. 2:** Comparison of observed and predicted yield using stepwise regression.

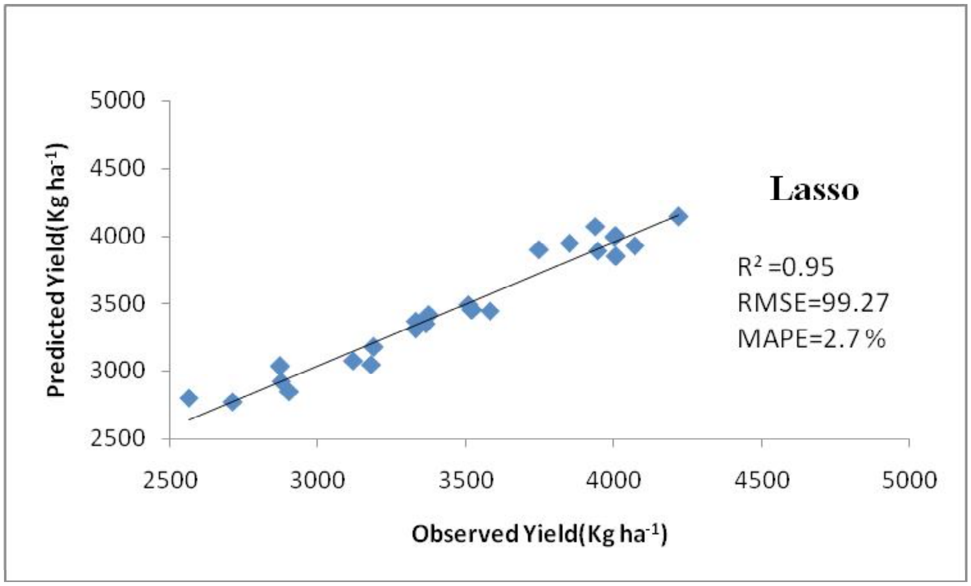

**Fig. 3:** Comparison of observed and predicted yield using lasso regression.

**Table 4:** Validation of forecast model with stepwise and Lasso regression during 2013-14 and 2014-15

| Year | Observed | Stepwise regression | | Lasso regression | |
|---|---|---|---|---|---|
| | yield | Predicted | Error (%) | Predicted | Error(%) |
| 2013-14 | 4134 | 3781 | -8.5 | 4211 | 1.89 |
| 2014-15 | 3905 | 4301 | 10.14 | 3969 | 1.64 |

variation between observed and simulate yield whereas Fig.3 shows less variation. Therefore, it can be inferred that the Lasso gives the best fit model as compared to stepwise regression method.

### *Forecast model validation*

Forecast models were validated in order to evaluate the performance of both stepwise and Lasso regression during 2013-14 and 2014-15. Results are presented in Table 4, prediction error with stepwise -8.5 per cent and 10.14 per cent and with Lasso 1.89 per cent and 1.64 per cent. It clearly implied that prediction accuracy of Lasso is better than stepwise at field level.

## CONCLUSIONS

Stepwise and Lasso regression variable selection were used to evaluate the performance of prediction accuracy in this study. Weather parameters were used to develop regression forecast model forty five days before harvest. Based on forecast model results, it was found that stepwise forecast model over fit, whereas Lasso performs better fit model. During validation period (2013-14 & 2014-15) also the per cent error by Lasso regression model was less than Stepwise regression. Therefore, it can be inferred that Lasso variable selection method performed better than stepwise at some extent.

## REFERENCES

Agrawal, R., Chandrahas and Aditya K. (2012). Use of discriminant function analysis for forecasting crop yield, *Mausam,* 36(3): 455-458.

Ghosh, K., Balasubramanian, R., Bandopadhyay,S., Chattopadhyay, N., Singh, K.K. and Rathore, L.S. (2014). Development of crop yield forecast models under FASAL- a case study of *kharif* rice in West Bengal. *J. Agrometeorol.,* 16(1):1-8.

Kutner, M., Nachtsheim, C., Neter, J. and Li, W. (2004). "Applied Linear Statistical Models". 5th edition. McGraw-Hill/ Irwin.

Lee, B.H. (2011). Forecasting wheat yield and quality conditional on weather information and Mathematical Methods for Digital Computers. New York: Wiley.

Rai, T. and Chandrahas (2000). Use of discriminant function of weather parameters for regression. *American Statistician*, 44: 214–217.

Tannura, M.A., S.H. Irwin, and D.L. Good (2008). A Review of the Literature on Regression Models of Weather, Technology, and Corn and Soybean Yields in the U.S. Marketing and Outlook Research Report 2008-02, Department of Agricultural and Consumer Economics, University of Illinois at Champaign-Urbana.

Tibshirani, R.J. (1996): Regression shrinkage and selection via the Lasso. *J. Roy. Stat. Soc.*, Series B, 58(1):267–28.

Tibshirani, R.J. (1997). The LASSO Method for Variable Selection in the Cox Model. *Stat. Med.*, 16, 385- 395

Walker, G.K. (1989). Model for operational forecasting of western Canada wheat yield.*Agric. For. Meteorol.,* 44, 339–351.

*Received : October 2018 ; Accepted : May 2019*