P: ISSN No. 2231-0045 RNI No. UPBIL/2012/55438 VOL.-7, ISSUE-2, November-2018

E: ISSN No. 2349-9435 Periodic Research

Development of Forecasting Model for Sugarcane Productivity using Multiple Linear Regression with Genetic Algorithm

R.S. Rajput
Assistant Professor,
Deptt. of Mathematics, Statistics and Computer Science,
College of Basic Sciences,
G. B. Pant University of Agriculture & Technology,
Pantnagar

Anjali Pant
Head,
Deptt. of Applied Science,
Govt. Polytechnic College,
Shaktifarm, Uttarakhand

Santosh Kumar
Technical Assistant,
Deptt. of Mathematics, Statistics and Computer Science,
College of Basic Sciences,
G. B. Pant University of Agriculture & Technology,
Pantnagar India

Abstract

We projected a Multiple Linear Regression forecasting model of productivity of sugarcane on the basis of data related to sugarcane productivity and weather parameters obtained from university farm , G.B. Pant University of Agriculture & Technology, Pantnagar, India (28.9700^o N, 79.4100^o E) . Then apply GA to improve MLR forecasting model by updating the value of Multiple Linear Regression coefficients, and selection of variables. Further again apply interaction of variables to improve the forecasting model. Comparison of developed models will be made by using indices like R^2, RMSE, Significance of dependent variables, Residuals, etc.

Keywords: Multiple Linear Regression, Genetic Algorithms, Forecasting, Sugarcane, R.

Introduction

Sugarcane crop

Sugarcane (*sacchaarum Officinarum*) is an important cash crop in the world (Takeo Yamane, 2018). The cultivation of sugarcane was extended to nearly all tropical and subtropical regions. Sugarcane growing countries of the world are lying between the latitude 36.7 degrees north and 31.0 degrees south extending from tropical to subtropical zones. It is long duration crop, and thus it encounters all the seasons viz. rainy, winter and summer during its life cycle. The sugarcane productivity and juice quality are profoundly influenced by weather conditions prevailing during the various crop-growth sub-periods (Amar Sawant, 2013).

Forecasting

It is a process of making statements about events whose actual outcomes had yet been observed. It is a branch of anticipatory science used for identifying and projecting alternatives possible future. Reliable and timely forecasts are of vital importance for appropriate foresighted and up-to-date planning in almost all the fields, especially for agriculture which is full uncertainties (Eurostat Statistics, 2014).

Multiple Linear Regression

A multiple linear regression (MLR) model that describes a dependent variable y by independent variables x_1, x_2, ..., x_m (m > 1) is expressed by the equation as follows, where the numbers α and β_k (k = 1, 2, ..., m) are the parameters, and ϵ is the error term (Kothari C.R. and Garg Gaurav, 2014). A dataset $\{y_i, x_{i1}, x_{i2}, \dots\dots, x_{im}\}_{i=1}^{n}$ of n statistical units, a multiple linear regression model assumes that the relationship between the dependent variable y_i and the m-vector of regressors x_i is linear. This relationship is modeled through a disturbance term or error variable ϵ_i an unobserved random variable that adds noise to the linear relationship between the dependent variable and regressors. Thus the model takes the form

$$y_i = \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots \dots + \beta_p x_{im} + \epsilon_i$$

or

$$y_i = X_i^T \beta + \epsilon_i \quad i = 1,2,\dots\dots n \tag{1}$$

where ^{T} denotes the transpose, that $x_i^T \beta$ is the inner product between vectors x_i and β.

$$y = X\beta + \epsilon \tag{2}$$

P: ISSN No. 2231-0045 RNI No. UPBIL/2012/55438 VOL.-7, ISSUE-2, November-2018
E: ISSN No. 2349-9435 Periodic Research

Where,

Dependent variable matrix

$$y = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix},$$

Regressors or independent variables

$$X = \begin{pmatrix} X_1^T \\ X_2^T \\ \vdots \\ X_n^T \end{pmatrix} = \begin{pmatrix} x_{11} & \cdot & x_{1m} \\ x_{21} & \cdot & x_{2m} \\ \vdots & \cdot & \cdot \\ x_{n1} & \cdot & x_{nm} \end{pmatrix}$$

Parameters

$$\beta = \begin{pmatrix} \beta_1 \\ \vdots \\ \beta_m \end{pmatrix},$$

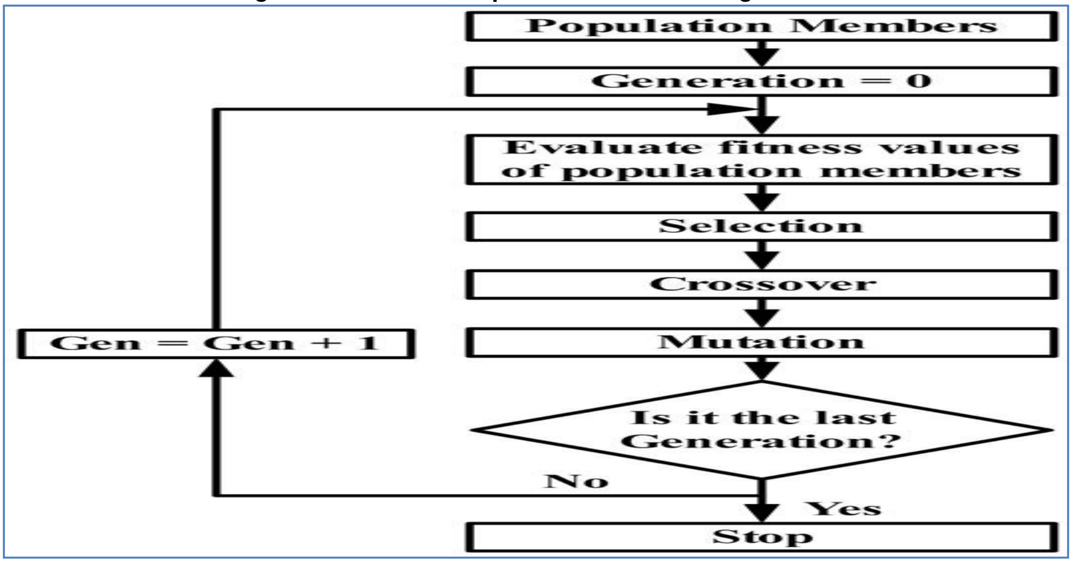
Error variables

$$\epsilon = \begin{pmatrix} \epsilon_1 \\ \vdots \\ \epsilon_m \end{pmatrix}$$

Genetic Algorithms

Genetic algorithms (GAs) are stochastic search algorithms inspired by the basic principles of biological evolution and natural selection. GAs simulates the evolution of living organisms, where the fittest individuals dominate over the weaker ones, by mimicking the biological mechanisms of evolution, such as selection, crossover, and mutation (Luca Scrucca, 2013)., (R. Leardi, 2009).

Figure-1 Flow-chart of process of Genetic Algorithms



An interaction occurs when an independent variable has a different effect on the outcome depending on the values of another independent variable. This is also known as a moderation effect.

Methodology

Objective of the Study

The objective of the study as following:

1. Development of MLR model for sugarcane productivity of Sugarcane.
2. Using a Genetic Algorithm to identify essential variables
3. Development of improved model using Genetic Algorithm
4. Development of improved model using the interaction of another variable

Data

In the present study we have also extended study of **(**Agrawal Ankuri, 2011) and data categorized two types as:-

1. Crop yield data and weather data. Yearly yield data of sugarcane (qt./ha) for 30 years (i.e., 1981 to 2011) were collected from University Farm.
2. Weather data of 30 years (from 1981 to 2011) were collected from Agro-meteorological Observatory of University.

We have selected 20 observations from the above data to build models.

Table 1: Notation of Parameters

Symbol	Parameter
T_1	Average Yearly Maximum Temperature (^{O}C)
T_2	Average Yearly Minimum Temperature (^{O}C)
H_1	Average Yearly Relative Humidity at 7.00 hrs
H_2	Average Yearly Relative Humidity at 14.00 hrs
R_f	Average Yearly Rainfall (mm)
W_s	Average Yearly Wind Speed (Km/h)
R_d	Average Yearly Number of rainy days
P_r	The productivity of Sugarcane (Qt/Ha)

Hypothesis

H_0: There are no significant relationships between P_r and (T_1, T_2, H_1, H_2, R_f, W_s, R_d)

H_A: There are some significant relationships between P_r and (T_1, T_2, H_1, H_2, R_f, W_s, R_d).

Model Testing parameters

t value

To test the null hypothesis, we compute a t-statistic, in the current paper we are using statistical software R. Accept null hypothesis if the probability of observing any value equal to | t | or larger. We test on the basis of a random if the mean of a population mean is the same as its hypothesized value of different (Kothari C.R. and Garg Gaurav, 2014).

Residual Standard Error (RSE)

Residual standard error of the estimate using following equation. (Gupta S.P. and Gupta M.P., 2009), (Kothari C.R. and Garg Gaurav, 2014)

$$RSE = \sqrt{\frac{1}{n-2}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \qquad (3)$$

The coefficient of determination (R^2)

The coefficient of determination (R^2) of a multiple linear regression model is the quotient of the variances of the fitted values and observed values of the dependent variable. (Gupta S.P. and Gupta M.P., 2009)

$$R^2 = \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \qquad (4)$$

Adjusted coefficient of determination (Adj R^2) : The adjusted coefficient of determination of a multiple linear regression model is defined in terms of the coefficient of determination as follows, where n is the number of observations in the data set, and m is the number of independent variables. (Gupta S.P. and Gupta M.P., 2009)

$$R^2_{adj} = 1 - (1 - R^2)\frac{n-1}{n-m-1} \qquad (5)$$

F-Statistics:_In MLR analysis F-test is used to test the overall validity of the model or to test if any of the explanatory variables are having a linear relationship with the response variable. (Kothari C.R. and Garg Gaurav, 2014)

Software used

R developed by R foundation for Statistical Computing. There is a *lm( )* function to perform Multiple Linear Regression (MLR) and, GA package of R use to perform GA computations.

Model Development

Step -1

Development of MLR model for sugarcane productivity of Sugarcane.

P_r=276.90582 + (-4.10810) * T_1 + (3.19319) * T_2 + (-2.19709) * H_1 + (0.29078) * H_2 + (0.02905) * R_f +(0.62845) * W_s + (1.27051) * R_d +ϵ

Summary of MLR model

```
Residuals:
Min          1Q          Median          3Q          Max
-5.4514      -1.8089     -0.7668         2.0680      6.2410
Coefficients:
             Estimate    Std. Error   t value    Pr(>|t|)
(Intercept)   276.90582  118.25248     2.342     0.0325 *
T1           -4.10810    2.29284      -1.792     0.0921 .
T2            3.19319    2.16943       1.472     0.1604
Rf            0.02905    0.02433       1.194     0.2500
Ws            0.62845    1.05941       0.593     0.5613
H1           -2.19709    0.92043      -2.387     0.0297 *
H2            0.29078    0.50320       0.578     0.5714
Rd           -1.27051    0.81599      -1.557 0.1390
Signif. codes: 0 '***'   0.001 '**'   0.01 '*'   0.05 '.'    0.1 ' '    1

Residual standard error: 3.938 on 16 degrees of freedom
Multiple R-squared:  0.5933,    Adjusted R-squared:  0.4154
F-statistic: 3.335 on 7 and 16 DF,  p-value: 0.02182
```

Residual standard error: 3.938 on 16 degrees of freedom

Multiple R-squared: 0.5933, Adjusted R-squared: 0.4154

F-statistic: 3.335 on 7 and 16 DF, p-value: 0.02182

Step -2

Genetic Algorithm, we have a set of 7 predictors (T_1, T_2, H_1, H_2, R_f, W_s, R_d) to predict P_r. We use GA to identify those predictors which are most relevant for explaining the variation of a response variable.

```
Residual standard error: 3.938 on 16 degrees of freedom
Multiple R-squared:  0.5933,    Adjusted R-squared:  0.4154
F-statistic: 3.335 on 7 and 16 DF,  p-value: 0.02182
GA settings:
Type=binary
Population size=50
Number of generations=100
Elitism =2
Crossover probability = 0.8
Mutation probability = 0.1
GA results:
Iterations = 100
Fitness function value = -102.1998
Solution =
       t1    t2    rf    ws    h1    h2    rd
[1,]   1     1     1     0     1     0     1
```

Step -3: Development of An Improved Model

P_r=206.26963 + (-3.42289) * T_1 + (3.02300) * T_2 + (-1.36542) * H_1 + (0.02516) * R_f + (-1.10980) * R_d

Summary of MLR model

Residuals:

Min	1Q	Median	3Q	Max
-3.6344	-1.6789	0.0543	1.4059	4.9034

Coefficients:

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	206.26963	81.94340	2.517	0.0246 *
T_1	-3.42289	1.41079	-2.426	0.0294 *
T_2	3.02300	1.24392	2.430	0.0291 *
R_f	0.02516	0.01696	1.484	0.1599
H_1	-1.36542	0.58363	-2.340	0.0346 *
R_d	-1.10980	0.53832	-2.062	0.0583 .

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.623 on 14 degrees of freedom
Multiple R-squared: 0.6517, Adjusted R-squared: 0.5274
F-statistic: 5.24 on 5 and 14 DF, p-value: 0.00642

Step -4: Development of improved model using the interaction of another variable

P_r=223.631515 + (-3.815362) * T_1 + (3.489946) * T_2 + (-1.306121) * H_1 + (-0.136790) * R_f + (-3.332454) * R_d + (0.019305)*rf * rd

Summary of the MLR Model

Residuals:

Min	1Q	Median	3Q	Max
-3.3714	-1.3227	-0.0367	1.2465	4.4944

Coefficients:

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	223.631515	75.316521	2.969	0.0109 *
t1	-3.815362	1.303234	-2.928	0.0118 *
t2	3.489946	1.160266	3.008	0.0101 *
h1	-1.306121	0.533539	-2.448	0.0293 *
rf	-0.136790	0.084439	-1.620	0.1292
rd	-3.332454	1.240637	-2.686	0.0187 *
rf:rd	0.019305	0.009895	1.951	0.0729 .

Signif. codes:0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.394 on 13 degrees of freedom
Multiple R-squared: 0.7306, Adjusted R-squared: 0.6063
F-statistic: 5.877 on 6 and 13 DF, p-value: 0.00372

Results and Discussion

Model 1

P_r=276.90582 + (-4.10810) * T_1 + (3.19319) * T_2 + (-2.19709) * H_1 + (0.29078) * H_2 + (0.02905) * R_f +(0.62845) * W_s + (1.27051) * R_d +ϵ

Model 2

P_r=206.26963 + (-3.42289) * T_1 + (3.02300) * T_2 + (-1.36542) * H_1 + (0.02516) * R_f + (-1.10980) * R_d

Model 3

P_r=223.631515 + (-3.815362) * T_1 + (3.489946) * T_2 + (-1.306121) * H_1 + (-0.136790) * R_f + (-3.332454) * R_{d+} (0.019305)*rf * rd

Comparison of models as given in the Table 2 based on parameters discussed in the previous section.

Table 2: Model comparison

Model constraint	Model No. 1	Model No. 2	Model No. 3
R^2	0.5933	0.6517	**0.7306**
Adjusted R^2	0.4154	0.5274	**0.6063**
Significance	T_1(p=0.0921) H_1(p=0.0297)	T_1(p=0.0294) T_2(p=0.0291) H_1(p=0.0346) R_d(p=0.0583)	T_1(p=0.0118) T_2(p=0.0101) H_1(p=0.0293) R_d(p=0.0187) R_f:R_d(p=0.0729)
Residuals	Min -5.4514 1Q -1.8089 Median-0.7668 3Q 2.0680 Max 6.2410	Min -3.6344 1Q -1.6789 Median0.0543 3Q 1.4059 Max 4.9034	Min -3.3714 1Q -3.3227 **Median -0.0367** 3Q 1.2465 Max 4.4944
Residual standard error	3.938	2.623	**2.394**
F-statistic	3.335	5.24	5.877
p-value	0.02182	0.00642	0.00372

Based on table 2, R^2 of Model No. 3 is highest than other Models. Adjusted R^2 of Model No 3 is again larger than other Models. A number of significant variables in Model No 3 are higher than the others. The residual standard error of the third Model is also less than the others. p-value third Model is also less than the others.

Conclusion

Based on results of study, we concluded Model No. 3 (P_r=223.631515 + (-3.815362) * T_1 + (3.489946) * T_2 + (-1.306121) * H_1 + (-0.136790) * R_f + (-3.332454) * R_{d+} (0.019305)*rf * rd) is best Model. Model No. 3 has highest R^2 value, minimum standard residual error. This Model is highly influence with the interaction of Average Yearly Rainfall (R_f) and Average Yearly Number of rainy days (R_d) and Average Yearly Maximum Temperature (T_1), Average Yearly Minimum Temperature (T_2), Average Yearly Relative Humidity at 7.00 hrs (H_1) and Average Yearly Number of rainy days (R_d) play significant role in the production of sugarcane.

References

1. *Agrawal Ankuri (2011). A comparative study of forecasting models. MS Thesis submitted to G.B. Pant University of Agriculture and Technology, Pantnagar, India.*
2. *Amar Sawant (2013), How to start Sugarcane Farming. Retrieved from https://agricultureguruji.com/start-sugarcane-farming/*
3. *Gupta, S.P. and Gupta, M.P. (2009). Business Statistics. Sultan Chand & Sons, New Delhi*
4. *Kothari, C.R. and Garg Gaurav (2014). Research Methodology Methods and Techniques Third Edition, New agre International limited, New Delhi*
5. *Luca Scrucca (2013) GA: A Package for Genetic Algorithms in R. Journal of Statistical Software, Volume 53, Issue 4. http://www.jstatsoft.org/*
6. *Takeo Yamane (2018) Sugarcane plant, Encyclopædia Britannica, Inc. retrieved from https://www.britannica.com/plant/sugarcane*
7. *Eurostat Statistics explained (2014) forecasting retrived from https://eceura.cu/eurostat/statistics-explained/idexphp/Glossary:Forecasting.*
8. *R Leardi(2009), Genetic Algorithms. Retrived from https://wwwsciencedirect.com/topics/medicine-and-dentistry/genetic-algorithms.*