Influential parameters on rainfall forecasting using multiple linear regression

Shobha N
Dept of Information Science and Engineering
A.P.S College of Engineering
Bengaluru

Dr. Asha T
Dept of Computer Science and Engineering
Bangalore Institute of Technology
Bengaluru

***Abstract*—In this study, the function of Multiple linear regression is used to find the influential parameters on rainfall forecasting. The algorithm is applied on the dataset to determine importance of input parameters on the target variable. The input parameters include humidity, temperature, cloud amount, wind speed and surface pressure. The regression model was evaluated statistically with R^2, adjusted R^2, p-value and residual standard error values.**

***Key words*- Multiple linear regression, Meteorological Data.**

I. INTRODUCTION

Prediction of Rainfall is one of the interrogations in meteorological observation. Rainfall is an important observatory of weather parameters for their application in weather and climate prediction. Weather affects agriculture at every step because of heavy rainfall due to variations in weather fronts like floods, cyclones etc and less rainfall as decreasing available moisture in atmosphere by cause of pollution. Knowledge in weather relationship helps in optimizing the agricultural production. Various studies of rainfall prediction have been addressed by many researchers, for example artificial neural network model [1, 2, 3] and multiple regression model [4, 5]. In this paper, influential parameters for rainfall prediction using multiple linear regression on climate observations is proposed. Performance of multiple linear regression is outlined in terms of coefficient of determination (R^2), F-statistic and p-values.
Section II briefly explained about related work. Section III describes on given methodology. Section IV concise on results and discussion followed by conclusion in section V.

II. RELATED WORK

Machine learning is a type of artificial intelligence, where machine can learn on its own by past experiences and discover new solutions. Artificial neural networks (ANNs), multiple linear regression (MLR), k-nearest neighbours (K-NN) and support vector machine (SVM) algorithms have contributed for prediction process.

Artificial neural network method was proposed to predict catchment flows in a snow dominated mountainous basin in Karasu Basin, Turkey by (Yilmaz et al.,2011) [1]. Twelve years meteorological data was used to calibrate ANN model and calibrated model was used to predict catchment flows. Model was evaluated based on linear regression and R^2.

To generate site-specific quantitative forecasts of daily rainfall, artificial neural network technique was used by (Maria et al.,2005) [2]. Meteorological variables such as potential temperature, vertical component of the wind, specific humidity, air temperature, precipitable water, relative vorticity and moisture divergence flux are used as input data to feed forward neural network and resilient propagation learning algorithm to predict rainfall forecast.

Factors that influence precipitation, were extracted from geographic variables of Jeju Island, Korea using multiple regression by (Myoung et al., 2011) [3]. The significance of the model was examined using F-test and t-test.

The application of Artificial Neural Networks (ANN) and Multiple regression analysis (MR) to forecast long-term seasonal spring rainfall in Victoria, Australia was investigated by (Mekanik et al., 2013) [4]. The ANN was implemented in the form of multilayer perceptron using Levenberg–Marquardt algorithm. Multilinear Regression and ANN modelling were evaluated statistically using mean square error (MSE), mean absolute error (MAE), Pearson correlation (r) and Willmott index of agreement (d).

III. METHODOLOGY

A. The dataset

The dataset comprise of meteorological observations gathered from Indian meteorological department, Bengaluru.

B. Multiple Linear Regression

Multiple linear regression is useful for modeling the relationship between a numeric or dependent variable (y) and multiple explanatory or independent variable (x). The scatter plot describes whether the variables are related to each other in linear way or not. The linear equation is given by

$$y = b_0 + b_1 x_1 + b_2 x_2 + \ldots . + b_n x_n$$

Where b_0 is an y intercept, $x_1, x_2, \ldots x_n$ are independent variable and y is dependent or predicted variable.

The independent variables in the data set were temperature, humidity, cloud amount, wind, surface pressure and dependent or target variable is rainfall. Steps to get regression results were shown below.

1. Multiple regression model is developed by calling the function linear model.
2. The intercept value and p-value shows how variables are significant in the model. Based on these value the variables which are not significant will be eliminated and apply the function linear model on the data set.
3. The value of intercept, p-value, multiple R-squared and adjusted R-squared for the modified model was calculated.

4. Fit the model by applying analysis of variance (ANOVA) to the significant variables in the data set.
5. Compute the intercept, p-value, multiple R-squared and adjusted R-squared values for ANOVA model.
6. Analysis of variance table was calculated between significant and not significant residuals by applying ANOVA model.

IV. EXPERIMENTAL RESULTS

R language is used to build regression model. We use lm and anova for multiple linear regression analysis.

TABLE I. COEFFICIENT VALUES FOR MODEL 1

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	97.148685	118.955566	0.817	0.414218
temp700	-0.246525	0.143195	-1.722	0.085313.
temp1400	0.106334	0.133512	0.796	0.425881
hum700	0.074729	0.035035	2.133	0.033059*
hum1400	0.094848	0.032994	2.875	0.004091**
cld700	0.839345	0.188144	4.461	8.65e-06***
cld1400	0.154420	0.252942	0.610	0.541611
wind700	0.015414	0.004465	3.452	0.000569***
wind1400	-0.002995	0.004751	-0.630	0.528512
slp1	0.250462	0.160153	1.564	0.118018
slp2	-0.369989	0.199378	-1.856	0.063657.

TABLE II. COEFFICIENT VALUES FOR MODEL 2

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-10.886786	2.881056	3.779	0.000163***
hum700	0.071280	0.034448	2.069	0.038665 *
hum1400	0.094848	0.032994	2.875	0.004091**
cld700	0.864965	0.151398	5.713	1.29e-08***
wind700	0.012443	0.002408	5.167	2.64e-07***

As shown in the Table I, the variable having 3 stars indicates that variable is highly significant, then p-value is equal to 0.001i.e (1-p) = 1-0.001 nearly equal to 0.999 or 99.9%. The variable having 2 stars and one star indicates that variable is less significant and contributes less to the model.
Less significant variables were removed and regression model was rebuilt, the results were tabulated in Table II. As shown in Table II all variables are significantly contributing towards the model 2. The intercept value obtained from this model is equal to -10.8867. The complete regression equation is written as

Rainfall= -10.8867+ 0.0712 * hum700 + 0.09484 * hum1400+ 0.8649*cld700+0.0124*wind700

As shown in the above equation, residuals such as humidity, cloud amount and wind perform significant role in rainfall event prediction.

TABLE III . VALIDATION VALUES FOR MODEL 1

Model	R square	Adjusted R square	Residual standard error
1	0.09946	0.0945	8.643

TABLE IV. VALIDATION VALUES FOR MODEL 2

Model	R square	Adjusted R square	Residual standard error
1	0.0962	0.09422	8.643

The evaluation values such as R^2, adjusted R^2 and residual error for model 1 and model 2 was demonstrated in Table III and Table IV.

Analysis of variance (anova) was applied on model 1 and model 2 and results were shown in the Fig 1, 2, 3 and 4.

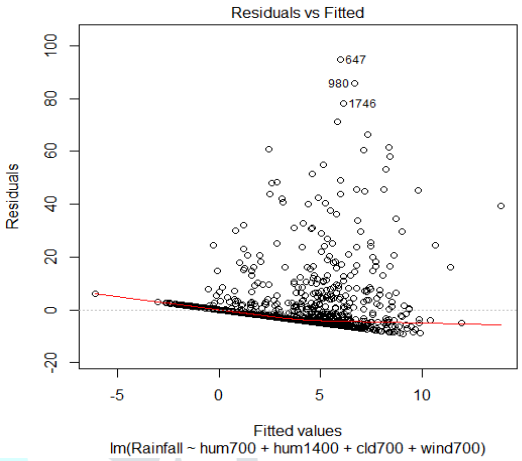

Fig 1 Residual vs Fitted

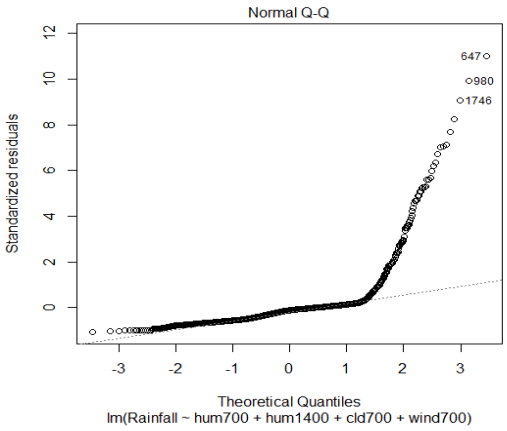

Fig 2 Normal Q-Q

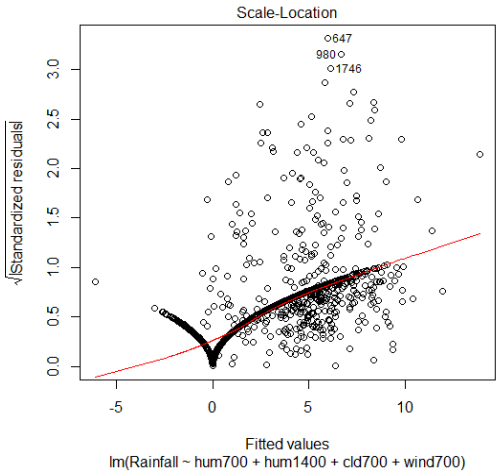

Fig 3 Scale-Location

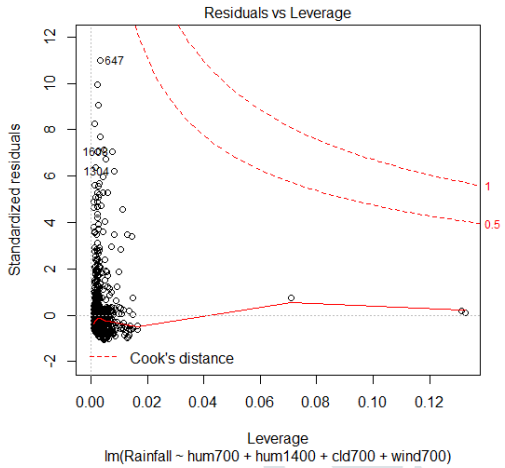

Fig 4 Residuals vs Leverage

In Fig 1, the residuals vs fitted graph shows some of the points lie above and below the plane and some points are outliers. It is a nonlinear fit which means small degree of linearity. In Fig 2, the normal Q-Q plot explains all the points lie on the line of significant model. In Fig 3, the scale location shows distribution of residuals on linear model against rainfall. Density of points follows the line when the rainfall is medium. Fig 4, residual versus leverage graph explains by eliminating extra values the model becomes linear.

V. CONCLUSION

Multiple linear regression application was developed on meteorological residuals to find most influential parameters that responsible for rainfall forecasting. We found that humidity, cloud and wind are most significant parameters to predict rainfall forecast and contributes more to the model. The attainment of the model was measured by R square, adjusted R square and residual standard error.

REFERENCES

[1] Yilmaz, A.G., Imteaz, M.A., Jenkins, G., "Catchment flow estimation using Artificial Neural Networks in the mountainous Euphrates Basin". Journal of Hydrology , 410 (1–2), (2011) 134–140.

[2] Maria Cleofe Valverde Ramirez, Haroldo Fraga de Campos Velho, Nelson Jesus Ferreira, "Artificial neural network technique for rainfall forecasting applied to the Sao Paulo region", Journal of Hydrology 301 (2005) 146–162.

[3] Myoung-Jin Um, Hyeseon Yun, Chang-Sam Jeong, Jun-Haeng Heo, "Factor analysis and multiple regression between topography and precipitation on Jeju Island, Korea", Journal of Hydrology 410 (2011) 189–203.

[4] F. Mekanik, M.A. Imteaz, S. Gato-Trinidad, A. Elmahdi, "Multiple regression and Artificial Neural Network for long-term rainfall forecasting using large scale climate modes", Journal of Hydrology 503 (2013) 11–21.

[5] Chiang, Y.M., Chang, F.J., 2009. "Integrating hydrometeorological information for rainfall–runoff modelling by artificial neural networks". Hydrological Processes 23 (11), (2009) 1650–1659.

[6] R. Shukla, K. Tripathi, A. Pandey, I. Das, "Prediction of Indian summer monsoon rainfall using nino indices: A neural network approach, Atmospheric Research 102 (1-2) (2011) 99-109.

[7] T. Mandal, V. Jothiprakash, "Short-term rainfall prediction using ANN and MT techniques ", ISH J. of Hydraulic Engineering 18 (1) (2012) 20-26.

[8] H. Spath, "Algorithm 39: Clusterwise linear regression", Computing 22 (1979) 367-373.

[9] W. DeSarbo, W. Cron, "A maximum likelihood methodology for clusterwise linear regression" , Journal of Classication, 5 (2) (1988) 249-282.

[10] S. Ganey, P. Smyth, "Trajectory clustering using mixtures of regression models" , in: S. Chaudhuri, D. Madigan (Eds.), Proceedings of the ACM Conference on Knowledge Discovery and Data Mining, New York, 1999, pp. 63-72.

[11] L. Garcia-Escudero, A. Gordaliza, A. Mayo-Iscar, R. San Martin, "Robust clusterwise linear regression through trimming", Computational Statistics and Data Analysis 54 (2010) 3057-3069.

[12] http://www.imd.gov.in

[13] http://www.uasbangalore.edu.in/index.php/research/agromoterology