Error Reduction in Data Prediction using Least Square Regression Method

Aaryan Shekhar Jha[1], Manoj Verma[2]

[1]*Integrated M.Sc. Student, Department of Mathematics, NIT Patna*
[2]*Research Scholar, Department of Civil Engineering, NIT Patna*

Abstract - *In this paper, least square regression method is modified by implementing the most fitted data from the prediction divided into three parts, first part is used for the prediction using least square regression method and some part of its data is fitted with the second part of original data to minimize the error. Also, the pattern of data fitting is traced out and further used for new modified model of least square regression method. The validation of proposed model is done by matching the predicted data with layer to use it for further prediction. In short, firstly data is the third part of original data, and the effectiveness of proposed model is calculated by finding mean absolute error, mean relative error, root mean square error. These errors are compared with the error found from original least square regression method. In the proposed model, least square regression method is considered as the base and it is refined by shifting the predicted value to the most fitted value of input dataset. Further introduced method is checked for its validity by taking some test data. The method shows the excellent results with minimizing errors and noises.*

Key Words: Prediction, Forecasting, Data, Least Square Regression Method

1. INTRODUCTION

TIn current scenario, data forecasting is one of the biggest challenges. It involves with the uncertainties, accuracy and error. Knowledge of data prediction techniques are mandatory for forecasting data and its trend. Data prediction techniques are largely used in medical sectors, to predict the health outcome due to therapeutic interventions or to recommend the medicine for its impact on the health. And the many other industries like automobile industries, market industries, civil engineering industries, aviation and many more. Prediction is used in business to identify risks and opportunities. A lot of method has already been developed for this purpose and many of them are reliable up to certain extent. Prediction models are need to be validated and updated for their results. Sometimes it is referred as model tuning. In machine learning, it is done by hidden layers. There are various techniques [1] for data prediction and forecasting such as, least square method, maximum likelihood method AI, machine learning etc [2]. Artificial Intelligence and Machine Learning is advanced version of tools required for data prediction. These method uses multilayer architectures to find the nature of data from lowest to highest level, and this structured data is used to predict huge amount of data [3]. Although, error is introduced by the model at the stage of prediction and it is required to be minimized.

There are several types of data, that are needed to be handled by prediction model. Data are widely classified as structured and unstructured data. They can be parametric or non-parametric in nature. For the handling of parametric data, Kalman filtering model can be used and for the non- parametric data k-Nearest Neighbour (k-NN) model and artificial neural network (ANN) are implemented.

Data is the core of all the predictive analysis [4]. Fig.1 shows the interrelation between the data and different methods to analyse. For the parametric data, auto regressive integrated moving average (ARIMA) is used for time series analysis [5]. ARIMA has been amended by the various authors. Whereas for non-parametric data advanced method is required due to stochastic and non-linear nature of data. Sasu [6] uses the k-NN method for the time series prediction and Hamed et al. [7] uses the k-NN parametric regression to find the multi interval data prediction model. Support Vector Regression (SVR) is one of the most popular method used for prediction in machine learning. ARIMA was used to obtain three relevant time series that were the basis of neural network (NN) in aggregation stage.

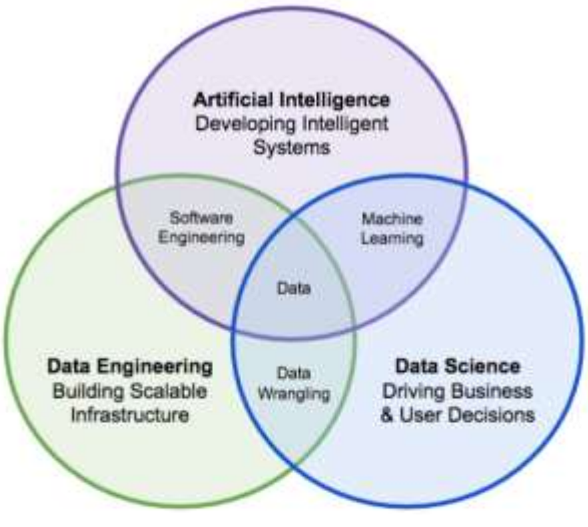

Fig -1: Relation between data

The existing methods are time and resource consuming and not perfect. They need to be refined. In this paper, least square regression method is modified by implementing the most fitted data from the prediction layer to use it for further prediction. In short, firstly data is divided into three parts, first part is used for the prediction using least square regression method and some part of its data is fitted with the second part of original data to minimize the error. Also, the pattern of data fitting is traced out and further used for new modified model of least square regression method.

The validation of proposed model is done by matching the predicted data with the third part of original data, and the effectiveness of proposed model is calculated by finding mean absolute error, mean relative error, root mean square error. These errors are compared with the error found from original least square regression method.

2. METHODOLOGY

In the proposed model, least square regression method is considered as the base and it is refined by shifting the predicted value to the most fitted value of input dataset. The methodology is shown in the Fig. 2 in form of flow chart. The given data set is divided into three parts. First part is used in least square regression method for the prediction of data and the part of predicted data is shifted to most fitted value of original dataset and fitting trends found out. This fitting trend can be further used to predict the future data.

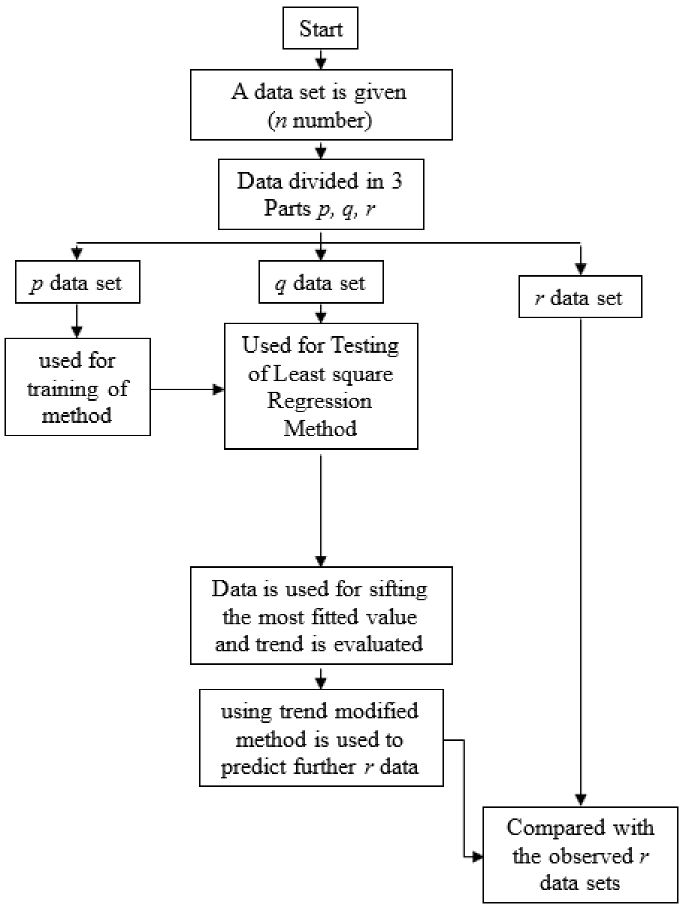

Fig -2: Flow chat of prediction model

2.1 Mathematical Formulation of proposed model

Let $(x, M(x))$ be the given dataset and suppose $(x, N(x))$ be the predicted dataset obtained using least square regression method [4], also let $(x, P(x))$ be the predicted dataset obtained from proposed model by shifting the most fitted value. Let there be n number of given input dataset. We divide the given dataset $(x, M(x))$ in three parts p, q and r such that $n = p + q + r$.

For the given input dataset $(x, M(x))$:
Let
$(x, M(x)) = \{(x_1, M(x_1)), (x_2, M(x_2)), \ldots\ldots\ldots, (x_p, M(x_p))\}$
then, $\overline{x} = \frac{x_1+x_2+x_3+\ldots\ldots+x_p}{p}$
and, $\overline{M(x)} = \frac{M(x_1)+M(x_2)+M(x_3)+\ldots\ldots+M(x_p)}{p}$,
where $\overline{x}$ and $\overline{M(x)}$ are respective mean values of x and $M(x)$ data.
Now, we know that to find the equation of a straight line, we need to find the slope and intercept. According to the least square regression method, the modified slope and intercept are given as,

$$m^* = \frac{\sum_{i=1}^{p}(x_i - \overline{x})(M(x_i) - \overline{M(x)})}{\sum_{i=1}^{p}(x_i - \overline{x})^2}$$

$$b^* = \overline{M(x)} - m^*\overline{x}$$

Therefore, the equation of the best fitted line obtained by least square regression method is given by

$$M(x) = m^*x + b^*$$

Now, using this equation, after putting each value of x data we would get our predicted model $(x, N(x))$. From this we get,
$N(x) = \{N(x_1), N(x_2), \ldots.., N(x_p), \ldots.., N(x_q), \ldots.., N(x_s)\}$, where $p < q < s$.
Now, p number of datasets is used to predict $p + q$ dataset whereas also s extra datasets are predicted in $(x, N(x))$.
$N'(x) = \{N(x_{p+1}), \ldots.., N(x_q), \ldots.., N(x_s)\}$ is used to shift the data to the most fitted value in original dataset $\{M(x_{p+1}), M(x_{p+2}), \ldots.., M(x_q)\}$. Such as for $(x_{p+1}, M(x_{p+1}))$, most fitted data is obtained from $(x, N'(x))$ as $(x_{p+a}, N(x_{p+a}))$ where a is distance of distance of data from p.
Similarly, this process is repeated for $\{M(x_{p+1}), M(x_{p+2}), \ldots.., M(x_q)\}$, and subsequent distances $u = \{a, b, c, d, \ldots..\}$ as the fitting trend dataset is obtained. This dataset is used to predict the minimum distance of next most fitted data point to predict the data.
Let the $u = \{a, b, c, d, \ldots\}$ is the most fitted data value minimum distance for the original $(x, M(x))$ dataset. This factor is used for modify the least square regression model $M(x) = m^*x + b^*$ as
$M(x) = m^*(x + u) + b^*$ such that the prediction model will become
$(x, N'(x)) = \{(x_{p+1}, N(x_{p+1})), \ldots.., (x_q, N(x_q)), \ldots.., (x_s, N(x_s))\}$ will become $(x, N''(x))$ which is predicted dataset from modified model.

2.2 Validation of model

The proposed model is validated for its least error in comparison to least square regression method by taking by mean absolute error (MAE), mean relative error (MRE), and root mean square error (RMS).

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|f_i - f_i^*|$$

$$MRE = \frac{1}{n}\sum_{i=1}^{n}\frac{|f_i - f_i^*|}{f_i}$$

$$RMSE = [\frac{1}{n}\sum_{i=1}^{n}(|f_i - f_i^*|)^2]^{1/2}$$

where, f_i = Observed data value
f_i^* = Predicted data value.

Table -1: Error Result and comparison

x-Range	Least square Regression Method			Proposed model			Remark
	MAE	MRE (%)	RMSE	MAE	MRE (%)	RMSE	
81-90	3.91	11.57	4.67	2.07	6.10	2.32	Less Error
91-100	1.91	5.48	2.84	1.78	4.97	2.43	Less Error
101-110	6.46	18.26	7.47	3.82	10.93	4.64	Less Error
111-120	1.49	3.82	2.10	1.69	4.38	2.55	More Error
121-130	1.31	2.86	1.48	1.26	2.71	1.68	Less Error
131-140	5.10	12.00	5.57	3.56	8.67	4.58	Less Error

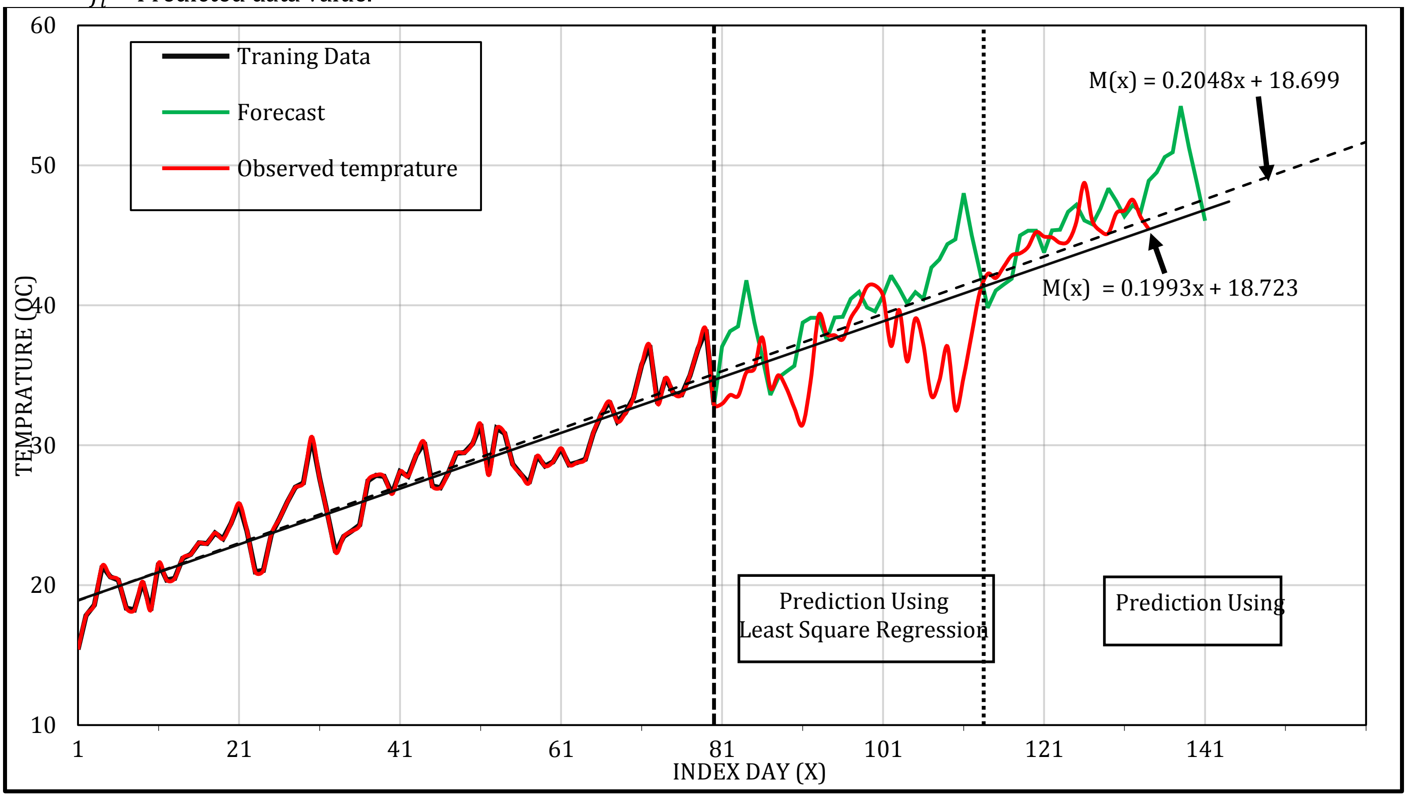

Fig -3: Graphical Representation of Analysis

For the validation of model, a test data is taken and predicted as per as proposed model. The maximum temperature data for 140 days is taken from Indian Meteorological Department (IMD) website. A simplified model is predicted for its tend using machine learning [8]. From Day1 to Day 80 data set is taken as the training data set denoted by p number of data in our proposed model. From Day 81 to Day 120 data is taken as testing data for least square regression method prediction denoted by q number of data. And from Day 121 to Day 140 the data is used for the prediction of error from both models.

3. RESULT AND DISCUSSION

The proposed model is obtained by modifying the least square regression method. And after validating the proposed model with different error estimation techniques, it can be observed that the errors obtained in the proposed model is significantly less than the error obtained from least square regression model. Fig. 4, Fig. 5 and Fig. 6 shows the comparison of errors that are obtained by modified model as well as least square regression method for MAE, MRE and RMSE errors respectively.

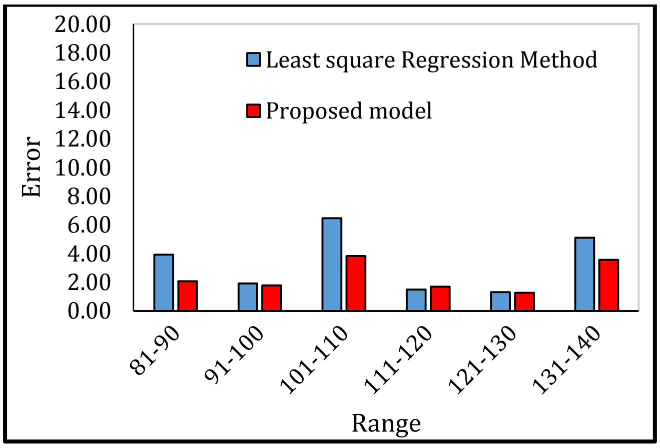

Chart -1: Comparison between Least square Regression Method and Proposed model for MAE

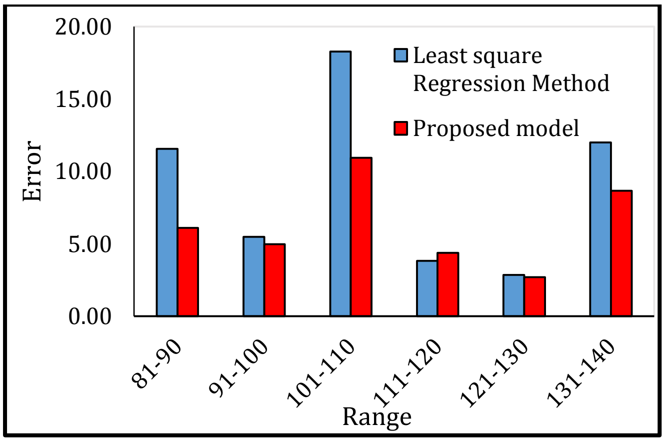

Chart -2: Comparison between Least square Regression Method and Proposed model for MRE

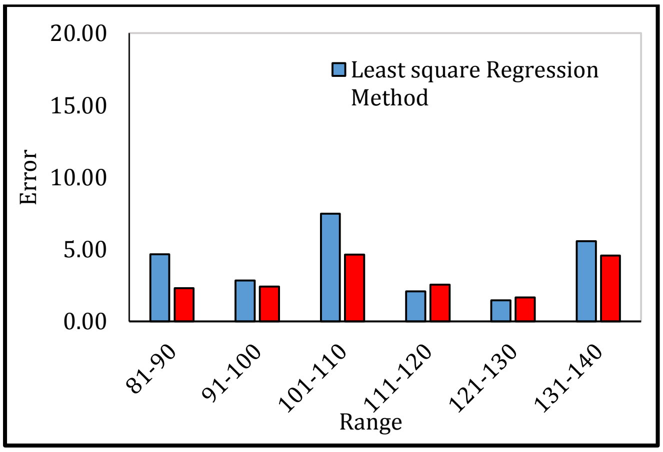

Chart -3: Comparison between Least square Regression Method and Proposed model for RMSE

4. CONCLUSION

In this paper, a model has been developed by modifying the least square regression method and validated by adopting numerical values to compute error. Model shows the less error as compared to the errors of least square regression method. The model has been focused on reducing the errors by altering the process of least square regression method. In this paper, a small sample dataset is used but this model can be checked for large number of data. This model is developed by focusing linear variable model, which comprises of one independent and one dependent variable. For future work, this model can be considered for multivariable and more complex data prediction. And it can be validated by using advanced machine learning techniques such as deep learning.

REFERENCES

[1] H. Seltam, "Experimental design and analysis," PsycCRITIQUES, vol. 20, p. 414, 2014.

[2] B. G. Subramaniam and T. R. Prabha, "Linear Regression in Machine Learning 1," vol. 2, no. 1, pp. 2–4, 2017.

[3] K. P. Moustris, P. T. Nastos, I. K. Larissi, and A. G. Paliatsos, "Application of multiple linear regression models and artificial neural networks on the surface ozone forecast in the greater Athens Area, Greece," Adv. Meteorol., vol. 2012, 2012.

[4] S. J. Miller, "The Method of Least Squares and Signal Analysis," pp. 1–7, 1992.

[5] A. A. Ariyo, A. O. Adewumi, and C. K. Ayo, "Stock Price Prediction Using the ARIMA Model," 2014 UKSim-AMSS 16th Int. Conf. Comput. Model. Simul., pp. 106–112, 2014.

[6] A. Sasu, "K-nearest Neighbor Algorithm for Univariate Time Series Prediction," Bull. Transilv. Univ. Brasov, vol. 5(54), no. 2, pp. 147–152, 2012.

[7] M. G. Hamed, M. Serrurier, N. Durand, M. G. Hamed, M. Serrurier, and N. Durand, "Simultaneous interval regression for K-nearest neighbor To cite this version: HAL Id: hal-00938894 Simultaneous Interval Regression for K -Nearest Neighbor," 2014.

[8] A. Agrawal, D. Verma, and S. Gupta, "Exploratory Data Analysis on Temperature Data of Indian States from 1800-2013," 2nd Int. Conf. Next Gener. Compuing Technol., vol. 2013, no. October, pp. 547–552, 2016.