The International Journal of Engineering and Science (IJES)
|| Volume || 11 || Issue || 2 || Series I || Pages || PP 08-12 || 2022 ||
ISSN (e): 2319-1813 ISSN (p): 20-24-1805

Piece Wise Isotonic and Simple Regression for Rainfall Data

V.N.Sendil[1], S.Venkatramana Reddy[2] and B. Sarojamma[3*]
1&3: Department of Statistics, S.V. University, Tirupati – 517 502, A.P., India.
2: Department of Physics, S.V. University, Tirupati – 517 502, A.P., India.

--***ABSTRACT***--
Now a days data mining techniques are rapidly growing in data science. There are many functions in data mining techniques (WEKA) and some of them are Gaussian process, Least Median Square, Linear Regression, Isotonic Regression, Pace regression, PLS Classifier, SMO regression, etc. In this paper we are fitting piece wise Isotonic Regression and Simple Regression models for Coastal Andhra Pradesh, Telangana and Rayalaseema regions for rainfall data from the year 1970 to 2017. We divide 1970 to 1989 as division I, 1990 to 2004 as division II and 2005 to 2017 as division III. Best model is chosen between isotonic regression and simple regression model using Root Mean Square Error criteria.

Keywords: *Annual Rainfall, piece wise isotonic regression, piece wise simple regression, Root Mean Square Error.*

Date of Submission: 05-02-2022 Date of Acceptance: 18-02-2022

I. INTRODUCTION

There are various Regression models in the literature and some of the important Regression models are Simple linear regression, Polynomial regression, Logistic regression, Ridge regression, Quantile regression, Lasso regression, Elastic net regression, Partial least squares regression, Principle components regression, Support vector regression, Ordinal regression, Poisson regression, Negative binomial regression, Quasi poisson regression, Cox regression, Piece wise linear regression, Probit regression etc. According to type of data we have to fit appropriate regression model. Generally for fitting regression models, we may make use of outliers, heteroscedasticity, multicollinearity, under fitting and over fitting.

Patrick Royston et al[1] published a paper entitled "Multivariate modelling with cubic regression splines: A Principle approach". In this paper they built multivariate fractional polynomials approach for building and goodness of fit was also tested for regression spline. Jerome H. Fried Man [2] explained MARS to pure noise function with three sample sizes (N=50, 100, 200) and two dimensionalities, n=5 and 10. Badi H. Baltagi et al[3] in their a paper "Unequally spaced panel data regression with AR(1) disturbances". they tested zero serial correlation in unequally spaced panels by Grunfeld Data. Simon N. Wood et al [4] is their research paper "Generalized Additive models with integrated model selection using penalized regression splines and applications to environmental modelling". They explained basic mathematical and numerical approach to Generalized Additive models in R package with examples of environmental. Enno Mammen and saravan de geer [5] in their paper on locally Adaptive Regression splines, Least Squares Splines with locally data adaptive placed knot points are considered and an iterative algorithm based on stepwise addition and deletion of knot points are proposed and their consistency proved.

Md.Naral Amin and Md. Ahsan Habib[6] published an article entitled "Comparison of different classification techniques using WEKA for Haematological dataset". They took data samples of 600 and analysed. Data set consists of 298 samples for a given CBC test. White blood cells, Red Blood cells, Hemoglobin Features using Haematorists Normal values of Males and Females separately. They took the parameters such as Mean Cellular Volume, Mean Cellular Hemoglobin, Mean cellular Hemoglobin concentration, Platelet count, Neutrophils, Lymphocytes, Monocytes, Eosinophils, Basophils. The classifiers used for study was Decision Tree(J48), Naïve Bayes and Multilayer Neural Network. The results says that J48 Decision tree was the best among 3 models taken using Kappa Statistic. The Relationship between Wind Speed and Precipitation in the Pacific ITCZ was given by Carissa E Back et al[7]. The paper contains data of 4 years passive microwave satellite retrievals from the SSMI and TMI used to look at the relationship between daily wind speed and Precipitation. Correlation between Wind speed and Precipitation was significant. The slope of relation between Wind speed and Precipitation was increased in moister conditions. The area averaged Precipitation estimates derived from a radar at Kwajaleivi Island compared with microwave Precipitation estimates 2.50 Vector Mean Winds computed from Quick SCAT with the SSMI-and TMI –derived Wind speeds. Vidyulatha Pellakure et al[8] has published a paper entitled "Applying Regression Techniques Environmental Data by WEKA". In this paper, they discussed Correlation, Regression and Prediction using Data mining process by WEKA tools for air pollutant data. "Machine Learning strategies for TimeSeries Forecasting"published by Gianluca Bontempi et al[9].

Generally piece wise regression is using simple regression for the different piece or sub divisions of the model. By using Knots we divide data in to different parts. In this paper we are fitted isotonic regression as well as simple regression models for rainfall data from 1970 to 2017[10]. Data is divided into 3 subdivisions first subdivision is from 1970 to 1989, second subdivision is from 1990 to 2004 and third sub division is from 2005 to 2017. Which model is the best between isotonic and simple regression is chosen by using root mean square error criteria.

II. METHODOLOGY

Generally piece wise linear regression is used for fitting regression lines for pieces of data of knots. Knots divided the data into pieces. For each piece we fit a separate regression line. In this paper, we are used Data mining techniques such as isotonic regression and simple regression models for different pieces of data. By using WEKA software, we are fitted isotonic regression and simple regression models.

Step by step procedure in WEKA for Isotonic Regression is as follows.
Step-1: Go to WEKA software and open
Step-2: WEKA explorer contains, pre-process as high light.
Step-3: In pre-process, bold keys are open file, open URL, open DB and generate.
Step-4: Go to open file, click on open file
Step-5: Choose appropriate file for analyse.
Step-6: Go to classify, in that choose WEKA from those functions.
Step-7: From functions choose isotonic regression
Step-8: Click on start, it generates output.

Step by step procedure for fitting simple regression using WEKA is as follows.
Step-1: Go to WEKA software and open
Step-2: WEKA explorer contains, pre-process as high light.
Step-3: In pre-process, bold keys are open file, open URL, open DB and generate.
Step-4: Go to open file, click on open file
Step-5: Choose appropriate file for analyse.
Step-6: Go to classify, in that choose WEKA from those functions.
Step-7: From functions choose linear regression
Step-8: Click on start, it generates output.
Various accuracy of measures used for Rainfall data are Mean Absolute Error, Root Mean Square Error, Relative Absolute Error and Root Relative Square Error.

III. EMPIRICAL INVESTIGATIONS

For Annual Rain fall data of coastal Andhra Pradesh, Telangana and Rayalaseema from the year 1970 to 2017 is splitted into three subdivisions by taking 2 knots. First sub division contains 1970 to 1989, second sub division from 1990 to 2004 and third sub division from 2005 to 2017. For each subdivision isotonic regression and simple linear regression models are fitted. The isotonic regression and simple regression models fitted values and actual values for three regions i.e., Coastal Andhra Pradesh, Telangana and Rayalaseema for subdivision-I are presented in Table-1

Table-1

Rayalaseema			**Coastal Andhra Pradesh**			**Telangana**		
Actual value	**Isotonic Estimate**	**Simple Regression**	**Actual value**	**Isotonic Estimate**	**Simple Regression**	**Actual value**	**Isotonic Estimate**	**Simple Regression**
826	802.6	758.7	1028.5	975.35	1015.289	670.7	936.45	1026.356
769.5	703	758.7	1079.9	1097.4	1015.289	976.9	1352.95	1026.356
817	803.6	763.85	1090.3	1012.11	1018.094	1035.6	1010.08	953.972
685.8	727.9	763.85	967.6	1012.11	1018.094	785.9	1056.125	1120.546
785.2	684.313	751.328	929.3	1004.889	1014.35	1143.5	949.767	993.299
943	684.313	751.328	1196	1004.889	1014.35	1497.5	949.767	1018.198

693.7	719.625	760.506	1030	1164.8	1017.517	1161	1544.9	1137.194
869.3	797.789	760.506	1038.3	1012.791	1017.517	1311.7	954.86	963.886
866.6	798.089	766.106	1132.8	1015.77	1020.128	808.6	1039.209	983.305
595.6	767.3	766.106	888.5	1015.77	1020.128	933.2	1039.209	1058.939
733	719.625	774.744	785.2	1057.767	1032.422	742.8	1066.9	1110.836
573.7	785.2	774.744	1014.8	1029.386	1032.422	922.6	1045.025	1039.28
813.4	726.1	759.017	934.1	1022.264	1021.406	791.7	1035.711	1002.198
776.4	807.45	759.017	1064.2	966.425	1021.406	970.3	699.9	850.41
728.5	831.143	768.561	944.9	1090.233	1032.2	830.4	1013.315	915.097
689.5	831.143	768.561	859.1	1090.233	1032.2	1042.5	790.033	899.085
580.9	741.013	780.561	1009.7	1079.9	1018.928	1007.3	1013.817	1021.422
621.1	826.863	780.561	1033.2	974.175	1018.928	729.1	820.5	1021.422
842.3	681.367	742.678	1164.8	974.48	1001.467	1544.9	1161	964.517
1041.6	778.644	742.678	1192.4	1013.08	1001.467	1215.8	929.275	964.517

The Isotonic regression and simple regression for sub division II by taking rainfall annual data as dependent variable and time from i.e.1990 to 2004 as independent variable using WEKA software. Estimated values and actual values for three regions i.e. Coastal Andhra Pradesh, Rayalaseema and Telangana are tabulated in Table-2

Table-2

Rayalaseema			**Coastal Andhra Pradesh**			**Telangana**		
Actual value	**Isotonic Estimate**	**Simple Regression**	**Actual value**	**Isotonic Estimate**	**Simple Regression**	**Actual value**	**Isotonic Estimate**	**Simple Regression**
596	887.1	889.728	actual	predicted,	1240.089	863.5	999.1	1100.441
1277.7	836.975	832.384	1001	1130.3	1114.57	986.8	974.5	1006.073
967.5	872.671	824.085	1210.9	1130.3	1267.187	849.1	1002.7	1118.206
741.7	670.35	824.085	1122.1	1124.514	868.065	667	864.25	852.841
615.2	732.225	841.215	873.6	900	984.684	765.3	916.433	862.998
871.3	732.225	841.215	703.2	977.067	1028.758	1078	916.433	901.867
887.1	859.625	814.392	992.3	977.067	1211.971	885.6	993.575	1055.877
948.1	777.667	814.392	869.7	1168.22	1045.514	1093	932.6	944.604
874.3	872.4	832.5	1310.8	942.233	1102.664	1286	939.056	975.25
725.5	678.45	832.5	1352	1086.6	855.997	963.2	667	812.905
924.5	694.133	821.271	1096.8	788.4	957.435	922.3	864.25	883.417
537.2	878.278	848.936	1009.3	900	1035.823	863.8	985.967	932.366
835	878.95	827.664	825.1	1000.8	1080.929	856	991.386	973.214
690.5	908.775	837.986	972.1	1145.943	1156.139	975.7	974.286	1028.493
930.7	967.5	820.829	1155.1	1119.417	1035.286	1425.7	974.463	932.521

Data mining techniques i.e. isotonic regression and simple regression are also used for estimating sub division III by taking Rainfall annual data as dependent variable and time i.e. from 2005 to 2017 as independent variable using WEKA software. Estimated values and actual values for three regions i.e. Coastal Andhra Pradesh, Rayalaseema and Telangana are listed in Table-3.

Table-3

Rayalaseema			Coastal Andhra Pradesh			Telangana		
Actual value	**Isotonic Estimate**	**Simple Regression**	**Actual value**	**Isotonic Estimate**	**Simple Regression**	**Actual value**	**Isotonic Estimate**	**Simple Regression**
963.5	752.94	785.245	1712.4	1076.467	1066.055	1276.4	980.44	961.918
738	752.94	785.245	861.9	1076.467	1047.563	753.1	980.44	961.918
638.1	1066.4	792.773	1159.6	1221.6	1074.564	1053	1163.1	974.027
980.6	816.567	792.773	1099.2	1203.467	1074.564	843.3	1035.3	974.027
551.8	816.56	839.273	874.9	1010.9	1117.745	746.4	857.3	983.709
555.4	816.56	839.273	908.9	892.7	1117.745	1043.4	815.7	983.709
715	774.875	802.017	1318.4	991.2	1063.383	1008.6	987.183	966.833
1066.4	638.1	772.733	1221.6	1173.84	1071.45	1163.1	1053	953.958
1047.1	647.233	683.072	1010.9	892.167	1089.008	857.3	894.9	979.442
798	817.6	795.1	1107.5	1190.425	1080.958	1035.3	982.733	964.608
762.6	762.975	798.05	1120.5	942.9	1079.875	1348.7	881.76	938.492
834.5	555.4	792.058	892.7	908.9	1098.858	815.7	882.367	982.908
688.2	845.05	804.25	790.5	1260.06	1164.029	666.3	1051.667	995.358

Table-4 explains the various measures of accuracy like Mean Absolute Error, Root Mean Square Error, Relative Absolute Error, and Root Relative Squared Error for Rayalaseema region, Coastal Andhra Pradesh region and Telangana region for sub division I, II and III.

Table-4

Measures of Accuracy	Rayalaseema			Coastal Andhra Pradesh			Telangana		
	Sub I	**Sub II**	**Sub III**	**Sub I**	**Sub II**	**Sub III**	**Sub I**	**Sub II**	**Sub III**
MAE	110.99	151.42	206.74	110.65	183.34	185.30	246.05	166.08	177.28
	103.64	150.39	160.07	88.89	189.99	194.11	215.21	169.02	183.09
RMSE	136.22	196.38	254.65	131.41	231.09	258.18	278.40	202.69	221.91
	127.43	190.09	195.98	112.88	240.31	254.30	263.52	206.58	214.04
RAE	107.09	103.32	135.09	124.47	95.63	96.93	117.75	109.95	96.83
	100	102.60	104.58	100	99.09	101.53	102.99	111.89	100
RRSE	106.89	105.34	138.23	116.41	96.15	102.33	107.18	100.41	103.67
	100	101.97	106.38	100	99.98	100.79	101.45	102.39	100

IV. SUMMARY AND CONCLUSIONS

Annual Rainfall data of united Andhra Pradesh from the year 1970 to 2017 is divided into three subdivisions by taking 2 knots. First sub division was from 1970 to 1989, second sub division was from 1990 to 2004 and third sub division was from 2005 to 2017. For Telangana region, Rayalaseema regions and Coastal Andhra Pradesh region we fitted isotonic regression and simple regression by WEKA. By using measures of accuracy like MAE, RMSE, RAE and RRSE, we explained which model is the best fitted models are listed in Table-5.

Table-5

Subdivisions	Best model Rayalaseema	Best model Coastal Andhra	Best model Telangana
1970-1989	Simple regression	Simple regression	Simple regression
1990-2004	Simple regression	Isotonic Regession	Isotonic Regession
2005-2017	Simple regression	Simple regression	Simple regression

REFERENCES:

[1]. Patrick Royston and williSauerbrej, "Multivariate modelling with cubic regression splines: A principled approach", The Strata Journal, vol. 7 no.1, (2007), pp 45-70
[2]. Jerome H. Fried man, "Annuals of Statistics", vol.19 no.1, (1991), pp1-67
[3]. Badi H. Baltagi and Ping X will, "Econometric theory", vol.15 no.6, (1999), pp814-823
[4]. Simon N. Wood and Nicok H. Augustion, "GAM with integrate model selection using penalized regression splines and applications to environmental modelling", Ecological modelling, vol.157 no2-3, (2002), pp157-177.
[5]. Enno Mammen and Sara Van De Geer, "Locally adaptive Regression splines", Annuals of Statistics, vol.25 no.1, (1997), pp387-413.
[6]. Md. Nurul Amin and Md. Ahsan Habib, "Comparison of different Classification Techniques using Weka for Hematological data", American Journal of Engneering Research, vol.4 no.3, (2015), pp55-61.
[7]. Larissa E.Back and Christopher S.Bretherton, "the relationship between Windspeed and Precipitation in the Pacific ITICZ", Journal of climate,vol.18 no.20,(2005), pp4317-4328.
[8]. Vidyullatha Pellakrui, D.Rajeswara Rao and Lakshmi Narayan, "Applying Regression Techniques on Environmental Data by WEKA", International Journal of Advances in Engineering and Research, vol.8, no.111,(2014), pp38-45.
[9]. Gianluca Bontempi, Souhaib Ben Taieb, and Yann-AeT Le Borgne, "Machine Learning Strategies for Timeseries Forecasting", eBISS Business Intelligence 2012, pp62-77, Springer-Verlag.
[10]. Data website:URL:http://www.imd.gov.in/Welcome%20To%20IMD/Welcome.php

V.N.Sendil, et. al. "Piece Wise Isotonic and Simple Regression for Rainfall Data." *The International Journal of Engineering and Science (IJES)*, 11(2), (2022): pp. 08-12.