Available online at www.ijpab.com

DOI: http://dx.doi.org/10.18782/2320-7051.6924

ISSN: 2320 – 7051

Int. J. Pure App. Biosci. **6 (5):** 90-96 (2018)

Research Article

International Journal of Pure & Applied Bioscience

OPEN ACCESS

Prediction of Rice Production in Madhya Pradesh through a Multiple Regression Approach

Kuldeep Rajpoot[1*], Abhishek Singh[2] and Thanu Ram Jaiswal[3]

[1,2]Department of Farm Engineering, BHU, Varanasi, (U.P.), India

[3]Department of Mathematics & Statistics, JNKVV, Jabalpur-482004, (M.P.) India

*Corresponding Author E-mail: kuldeep2rajpoot@gmail.com

Received: 25.08.2018 | Revised: 21.09.2018 | Accepted: 28.09.2018

ABSTRACT

During the last few decades, the statisticians, economists and other scientists have given due consideration to see the performance of the production of rice crop based on area under cultivation, cost of labour, cost of seed, cost of fertilizer etc. In the present study the multiple regression model has been fitted using least square principle. The test for normality of errors, homogeneity of error variances and independence of serial correlation of error terms (no autocorrelation) in the model have been investigated.

Key words: *Multiple regression, Autocorrelation, Durbin-Watson'd' statistic, Partial regression coefficients, Response and predictor variable, Rank correlation, J-B Test.*

INTRODUCTION

Rice *(Oryza sativa)* is important staple food crops grown in all parts of India.FThe area and production of Rice in India is 44.11 million hectares and 105.48 million tonnes, respectively with yield of 2391 kg/ha. Madhya Pradesh is also one of the major rice producing state in India, area and production of rice in Madhya Pradesh is 2.15 million hectares and 3.63 million tonnes, respectively with yield of 1684 kg/ha (Source Agricultural Statistics at a glance 2016).

The area, production and productivity level of this crop are varying from year to year. There is also a wide gap between the yield obtained on farmer's field and their potential. The present study is undertaken to see the status of area, production and production levels of Rice crop in Madhya Pradesh. Now the task of a statistician is to establish the actual relationship between response and predictor variables under study.

One variable (Response y) is related to various other variables (predictors x), many of which may interact among themselves Mulltiple Regression model may describe the production pattern, projected production and effect of diffrent input variables like area, fertilizer, seed, labour, price and level of technologies adopted under consideration.

Cite this article: Rajpoot, K., Singh, A., and Jaiswal, T. R., Prediction of Rice Production in Madhya Pradesh through a Multiple Regression Approach, *Int. J. Pure App. Biosci.* **6(5):** 90-96 (2018). doi: http://dx.doi.org/10.18782/2320-7051.6924

Mulltiple regression model Should follows some statistical assumptions like normality of error, homogeneity of error variances and independence of error terms (no autocorrelation) in the model. So in our study, we have tested these assumptions through different test statistics.

MATERIAL AND METHOD

Secondary data of wheat crop viz. production, yield, area under cultivation (X_1), price of seed (X_2) per quintal Production, price of manure & fertilizer (X_3) per quintal Production, fixed cost (X_4) per quintal Production, price of labour (X_5) per quintal Production in Madhya Pradesh state for 23 years (from 1990-91 to 2012-13) have been collected with the various sources. The collected data were compiled and analyzed in the view of objectives of the study under consideration.

Fitting of Multiple Regression Model

The postulated model is written by Damodar N. Gujrati (2004) as –

$$Y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{3i} + - - - - + \beta_5 x_{5i} + e_i \quad \ldots\ldots \quad (1)$$

Where β_0 ,β_j (j =1, 2,….,5)are unknown parameters to be estimated and e_i is the error term distributed as $e_i \sim N(0, \sigma_e^2)$ and (i = 1, 2, …..n), n is the number of years considered in the study. Using the principle of least square technique, the fitted multiple regression model is written as

$$\hat{y}_i = b_0 + \sum_{j=1}^{5} b_j x_{ji} \quad \ldots\ldots\ldots \quad (2)$$

where; "b_0" is the intercept term and b_j (j =1, 2, …., 5) are the estimate of partial regression coefficients β_j (j =1, 2,….,5) respectively. Then an estimate of y_i will be obtained as:

$$\hat{y}_i = \bar{y}_i + \sum_{j=1}^{5} b_j x_{ji} \quad \ldots\ldots\ldots \quad (3)$$

The multiple Correlation Coefficient (R) between response values (observed) y_i and its estimated values $\widehat{y_i}$ is worked out as

$$R^2 = \frac{\mathrm{cov}(y_i, \hat{y}_i)}{\sqrt{v(y_i), v(\hat{y}_i)}} \quad \ldots\ldots\ldots \quad (4)$$

We use the F statistics to test the significance of multiple Correlation Coefficient (R) written Kuswahwaha and Kumar[5] as

$$F_{(m, n-m-1,)} = \frac{R^2(n-m-1)}{(1-R^2)m} \quad \ldots\ldots\ldots \quad (5)$$

with statistical hypothesis stated as

Ho: R=0 VS H_1: R≠ 0

where m is the number of predictor variables considered in this study (here m=5).

Testing For Normality of Error Terms in Fitted Model

(i) JB Test for Normality

To test normality of error terms, the test statistics known as Jarqua-Bera (JB) test statistic (1985) is written as

$$J.B. = n\left[\frac{s^2}{6} + \frac{(k-3)^2}{24}\right] \quad \ldots\ldots\ldots \quad (6)$$

with statistical hypothesis stated as

Ho: Errors are normally distributed VS H_1: Errors are not normally distributed

where (S,K) are the skewness and kurtosis of errors. The JB statistic fallows asymptotically a chi-squared distribution, with 2 degrees of freedom. If the JB test statistic equals zero, it depicts that the error term (e_i) is normally distributed other wise not normally distributed.

(ii) Normal probability plot

A normal probability plot observed cumulative probabilities of occurrence of the standardized residuals on the Y axis and of expected normal probabilities of occurrence on the X axis, such that a 45-degree straight line will appear when the observed conforms to the normally expected and the assumption of normally distributed error was accepted.

Test for Homogeneity of Error Variances

Spearman's rank correlation coefficient defined as:

$$r_{s = 1-6}\left[\frac{\sum di^2}{(n^2-1)}\right] \quad \ldots\ldots\ldots \quad (7)$$

Where d_i is the difference in the ranks assigned to two different characteristics of the ith individual, n is the sample size. Assuming that the population rank correlation coefficient ρ_s is zero and n > 8, the significance of the sample r_s can be tested using student's t statistic defined as

$$t = \frac{|r_s|\sqrt{n-2}}{\sqrt{1-r_s^2}} \sim t(n-2). \quad \ldots\ldots \quad (8)$$

with statistical hypothesis stated as

Ho: Errors variances are homogeneous VS

H_1: Errors variances are not homogeneous

If the computed t value exceeds the critical t value, we may reject the hypothesis of homogeneity of errors variance, otherwise we may accept it. If the regression model involves more than one predictor variable X, r_s is computed between $|e_i|$ and each of the X

variables separately and can be tested for homogeneity of errors variances accordingly.

Testing For Serial Correlation of Error Terms in Fitted Model

In some cases, the use of an estimate of the serial correlation parameter indicate that the least square residuals may give even less efficient estimates than the original ordinary least square estimates. Therefore, researcher should be cautious in drawing inferences about the nature of serial correlation in the errors

Durbin- Watson has proposed[3] a widely used test for serial correlation in error terms based on the least square residuals. The test statistic is known as "Durbin –Watson d Statistic" which is defined as

$$d = \frac{\sum_{i=2}^{n} (e_i - e_{i-1})^2}{\sum_{i=1}^{n} e_i^2} \quad \text{.....} \quad (9)$$

with statistical hypothesis stated as

H0: Errors terms e_i's in fitted regression model are serially independent (no autocorrelation) i.e. $\rho_s = 0$ VS

H_1: Errors terms e_i's in fitted regression model are serially dependent (autocorrelation) i.e. $\rho_s \neq 0$

In the table for the given sample size n and number of parameter estimated (say m), the two critical values known as lower critical values d_L and upper critical values d_U are given and we conclude about the null hypothesis as fallows –

i. When computed value $d < d_L$ or $(4-d) < d_L$ H_0 may be declared significant.
ii. When computed value $d > d_U$ or $(4-d) > d_U$ then H_0 may be declared non significant.
iii. When $d_L < d < d_U$, then the test is inconclusive.

RESULTS AND DISCUSSION

The result of the present study for describing fitted multiple regression model of rice crop production, assumptions related to predicted errors like normality of errors, homogeneity of errors variances and independence of serial correlation error are described in this section as fallows.

Fitted Multiple Regression Model

The fitted multiple regression line has been obtained as:

$$y_i = -669.04 + 1.27\ x_{1i} + 6.88x_{2i} - 12.91x_{3i} - 1.56x_{4i} + 3.05x_{5i} \quad \text{.....} \quad (10)$$

The value of coefficient of multiple determination for rice crop has been obtained as:

$R^2 = 0.5488$

We use F test for significance of an observed coefficient of multiple correlation coefficient, for testing the hypothesis, the test statics F is computed, the calculated value of F statics given as:

Calculated $F_{(5, 17)} = 4.136$

and Tabulated $F_{0.05}(5,17) = 2.18$

Thus, we see that the calculated value of F statistic is much greater than the tabulated value which gives the evidence for the null hypothesis H_0 to be declared significant at 5% level of significance. Hence one can say that multiple correlation coefficients in population is high. In other words, one can finally conclude that the estimated value of production of rice as obtained by the multiple regression of production on the corresponding area under cultivation, cost of seed, Fertilizer, Human labour and artificial labour consumed for per quintal Production and actually observed value of production are highly correlated (closely associated). Hence a very accurate value of Production of rice can be estimated using the multiple regression equation of production of rice crop on the corresponding area under cultivation, yield, cost of seed, Fertilizer, labour and artificial labour consumed for per quintal Production.

The value of coefficient of multiple determination (R^2) = 0.5488 indicates that 54.88% of variation in the wheat production can be explained by the fitted multiple regression line for wheat crop.

Testing for Normality of Error Terms in Fitted Model

(i) JB Test for Normality

The fitted multiple regression line has been obtained as:

$$Y_i = -669.04 + 1.27\ x_{1i} + 6.88x_{2i} - 12.91x_{3i} - 1.56x_{4i} + 3.05x_{5i} \quad \text{.....} \quad (11)$$

From the fitted model we calculate estimated values $\hat{y_i}$, error $e_i = (y_i - \hat{y_i})$, the skewness (S) and kurtosis(K) for errors and are presented in the table 1.

Table 1: JB Test to Test the Normality of Errors Term

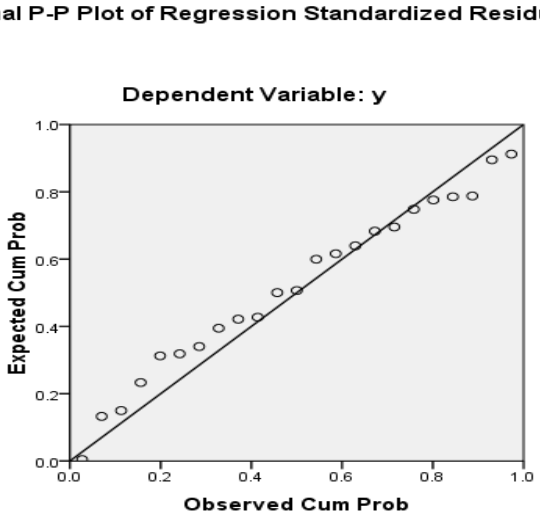
year	Area (x1)	Seeds (x2)	Fertilizer (x3)	H. labour (x4)	A.labour (x5)	Prod (y)	$\hat{y}$	ERROR (ei)
1991	1556	72.8	49.48	200.7	163	1435	1357.51335	77.486653
1992	1559	78.3	53.47	209.11	157	1543	1316.21976	226.78024
1993	1572	83.8	58.67	211.8	173	1165	1348.0138	-183.0138
1994	1566	90.13	58.89	235.4	186	1346	1383.72668	-37.726682
1995	1612	95.53	60.45	250.6	201	1459	1481.22913	-22.229125
1996	1672	97.2	61.67	286.39	224	1212	1567.43739	-355.43739
1997	1644	100.73	63.08	444.77	204	1346	1229.00615	116.99385
1998	1672	99.17	64.68	397.57	203	1424	1304.13008	119.86992
1999	1740	103.56	66.58	397.32	274	1750	1613.08806	136.91194
2000	1708	100.14	67.9	577.97	207	982	1044.89284	-62.892837
2001	1776	98.43	54.35	437.71	240	1692	1614.74052	77.259483
2002	1681	92.5	43.67	563.3	181	1032	1214.48272	-182.48272
2003	1719	96.35	46.89	504.44	240	1750	1519.66262	230.33738
2004	1686	100.01	50.34	510.7	263	1309	1518.52567	-209.52567
2005	1711	105.67	52.99	517.24	293	1694	1636.24191	57.758093
2006	1684	114.41	58.59	525.24	262	1396	1482.72404	-86.724037
2007	1645	124.07	63.14	529.36	221	1332	1309.41381	22.586185
2008	1717	130.4	64.89	538.06	264	1578	1539.44692	38.553078
2009	1446	132.45	65.5	547.18	284	1261	1247.09431	13.905687
2010	1470	135.89	66.67	515.78	300	1432	1384.11538	47.884618
2011	1500	137.56	71.24	560.78	342	1456	1432.32281	23.677185
2012	1507	140.33	76.47	562.23	385	1500	1521.49424	-21.494235
2013	1534	144.17	79.09	577.34	399	1539	1567.47466	-28.474663

Skewness = -0.67554
and kurtosis = 0.81676
By putting this value skewness (S) and kurtosis (K) in JB statistic we get
JB = 6.317272

For wheat crop calculated value JB statistics 6.317272and probability obtaining such a statistic under normality assumption is about 0.042484, which is very low which indicates that, we may reject the normality assumption under Jarqua-Bera (JB) test but graphical method for detection of normality i.e. Histogram of residuals and Normal probability plot indicates that error terms "follows normality assumption" here we used normal probability plot to test the normality assumption. The sample size n=23 <30 which is not sufficient for Jarqua-Bera (JB) test to fallow asymptotic normality of error terms that's why JB test rejecting the normality assumption.

(ii) Normal probability plot

By the JB test of normality of errors we get the result that error terms are not normally distributed but normality of errors is one of major assumption of regression model. The reason behind the rejection of normality assumption by JB test is that the it is asymptotic test which work well with the large size data where number of element should be greater than 30 but in our study we have take only 23 years data i.e. numbers of elements are 30 which is less than 30 so in that case we use graphical methods like normal probability plot or histogram of residuals which also suitable small size data. So here we are plotting normal probability plot to check normality of errors in fitted multiple regression model.



Graph 1: Normal probability plot in wheat

We can easily see that fitted line in the NPP is approximately a straight line. Hence one can say errors in fitted multiple regression model was normally distributed with mean approximately zero and standard deviation 0.69.hence one can finally say that the errors in fitted multiple regression model follows normality assumption for wheat.

Test for Homogeneity of Error Variance

Form the regression line of y of area (X_1), we calculate $e_i = (y_i - \hat{y_1})$ and ignoring the sign of e_i, the ranks are allotted to both e_i and X_{1i} according to either ascending or descending order of their magnitudes. These values are recorded in the table 2.

Table 2: Values of X_{1i}, e_i and Their Rank

year	Area(x1)	Error (e_i)	Rank(x_1)	Rank(e_i)	d_i(difference)	d_i^2
1991	1556	77.486653	6	20	14	196
1992	1559	17.003817	7	6	-1	1
1993	1572	65.893875	9	18	9	81
1994	1566	29.956286	8	10	2	4
1995	1612	12.582939	10	4	-6	36
1996	1672	50.727358	13	17	4	16
1997	1644	167.15117	11	23	12	144
1998	1672	18.747866	13	7	-6	36
1999	1740	79.722038	22	21	-1	1
2000	1708	11.653771	18	3	-15	225
2001	1776	30.253152	23	11	-12	144
2002	1681	31.077419	15	12	-3	9
2003	1719	85.988734	21	22	1	1
2004	1686	16.049811	17	5	-12	144
2005	1711	32.498337	19	13	-6	36
2006	1684	46.280995	16	15	-1	1
2007	1645	26.187864	12	9	-3	9
2008	1717	22.687225	20	8	-12	144
2009	1446	48.163144	1	16	15	225
2010	1470	42.451062	2	14	12	144
2011	1500	0.0503576	3	1	-2	4
2012	1507	1.0698627	4	2	-2	4
2013	1534	70.800798	5	19	14	196
Sum			275	276	1	1801

From the table Spearman's rank correlation coefficient is obtained as

$$r_{s = 1-6}\left[\frac{1801}{(23^2-1)}\right] = 0.1102$$

and calculated value of student's t statistic is obtained as

$$t = \frac{|0.1102|\sqrt{23-2}}{\sqrt{1-0.1102^2}} = 0.5079$$

Spearman's rank correlation coefficient r_s = 0.1102 and calculated value of student's t = 0.5079 where tabulated value $t_{(0.05,\ 21)}$ = 2.0796, which indicates that error H_0 is non significant and it will be accepted H_0 i.e. error variances are homogeneous in fitted regression line of y on X_1. Similarly we can test for homogeneity of error variances in fitted regression lines of y on other predictor variables (say X_2, X_3, X_4 and X_5) and we find out for each predictor variable error variances are homogeneous.

Testing for Serial Correlation of Error Terms in Fitted Model

Through the fitted regression model, the values of $\hat{y_1}$, $e_i = (y_i - \hat{y_1})$, e_i^2 and $\mathbf{(e_i - e_{i-1})^2}$ have been worked out and are given in the table (3) as bellow.

Table 3: Durbin-Watson d statistics for production of rice crop

year	prod(y)	$\hat{y}$	error (e_i)	($e_i - e_{i-1}$)	$(e_i)^2$	$(e_i - e_{i-1})^2$
1991	1435	1357.51335	77.486653		6004.181	
1992	1543	1316.21976	226.78024	149.2936	51429.28	22288.57435
1993	1165	1348.0138	-183.0138	-409.794	33494.05	167931.1536
1994	1346	1383.72668	-37.726682	145.2871	1423.303	21108.34688
1995	1459	1481.22913	-22.229125	15.49756	494.134	240.1742782
1996	1212	1567.43739	-355.43739	-333.208	126335.7	111027.7466
1997	1346	1229.00615	116.99385	472.4312	13687.56	223191.2727
1998	1424	1304.13008	119.86992	2.876077	14368.8	8.271819165
1999	1750	1613.08806	136.91194	17.04202	18744.88	290.430397
2000	982	1044.89284	-62.892837	-199.805	3955.509	39921.95018
2001	1692	1614.74052	77.259483	140.1523	5969.028	19642.67277
2002	1032	1214.48272	-182.48272	-259.742	33299.94	67466.01421
2003	1750	1519.66262	230.33738	412.8201	53055.31	170420.4398
2004	1309	1518.52567	-209.52567	-439.863	43901.01	193479.5081
2005	1694	1636.24191	57.758093	267.2838	3335.997	71440.61207
2006	1396	1482.72404	-86.724037	-144.482	7521.059	20875.0857
2007	1332	1309.41381	22.586185	109.3102	510.1358	11948.72461
2008	1578	1539.44692	38.553078	15.96689	1486.34	254.9416554
2009	1261	1247.09431	13.905687	-24.6474	193.3681	607.4938776
2010	1432	1384.11538	47.884618	33.97893	2292.937	1154.567756
2011	1456	1432.32281	23.677185	-24.2074	560.6091	585.9997796
2012	1500	1521.49424	-21.494235	-45.1714	462.0022	2040.457264
2013	1539	1567.47466	-28.474663	-6.98043	810.8064	48.72637096
sum	32633	32632.9969	0.0031486	-105.961	423336	1145973.165

Following Kuswahwaha and Kumar[5], computed value of 'd' statistics is obtain as

d = 1145973.16/423336 =2.707

where, m= 5, n=23 the table value at 5% level of significance for one tail test is given as:

d_L = 0.89` and d_u = 1.92,

Thus, we see that the computed d statistic is equal to 2.56 which is larger than d_u (1.92) hence the null hypothesis is declared non-significant. Hence we finally conclude that the errors involved in the fitted multiple regression model for rice crop are serially independent.

CONCLUSION

On the basis of statistical result obtained, the fallowing conclusions are drawn in the preset study. The multiple regression analysis reveals that the entire five predictor variables are found to be important characters for improving the production of Rice crop. In fitted multiple regression model for rice crop, error term follows normality assumption and Rank correlation test for homogeneity of error variances are found non-significant for all five input factors under cultivation i.e. errors term shows homogeneity in its variances. Durbin-Watson'd' statistic is found non-significant for rice crop which indicates that error terms are serially uncorrelated (no autocorrelation) in the fitted multiple regression model.

REFERENCES

1. Bajpai, R. K., Upadhyay, S. K., Joshi, B. S. and Tripathi, R. S., Productivity and economics of rice (Oryza sativa L.)-wheat (Triticum aestivum) cropping system under integrated nutrient supply systems, *Indian Journal of Agronomy*. **47(1):** 20-25, New Delhi (2002).
2. Damodar, N., Gujrati, Basic Econometrics, 4th ed., Mcgraw-Hill, New York (2004).
3. Durbin, J., Watson, G. S., Testing for serial correlation in least square regression. *Biometrika*, **38:** 159-171 (1951).
4. Jarque and Bera, A., Test for Normality of Observations and Regression *Residuals International Statistical Review / Revue Internationale de Statistique*, **55(2):** (Aug., 1987), pp. 163-172 (1985).
5. Kuswahwaha, K. S. and Rajesh, K., The Theory Of Sample Surveys And Statistical Decisions, first edition, New India Publishing Agency, New Delhi (2009).
6. Neeraj, k., Pisal, R. R., Shukla, S. P., Pandey, K. K., Crop Yield Forecasting of Paddy, Sugarcane and Wheat through Linear Regression Technique for South Gujarat. *Mausam;* **65(3):** Jul 2014; PP: 361-364 (2014).
7. Rajpoot, K. and Kuswahwaha, K. S., Prediction of Mustard Production in Madhya Pradesh Through A Multiple Regression Approach, *Trends in Biosciences* **10(22):** *Print*: *ISSN 0974-8431, 4500-4506*, (2017).
8. Sharma, V. P. and Joshi, P. K., Performance of rice production and factor affecting acreage under rice in costal regions of India. *Indian Journal of Agril. Econo.* **50(2):** 153-167, New Delhi (1995).
9. Wahab, N. S., Rusiman, M. S., Mohamad, M., Azmi, N. A., Him, N. C., Kamardan, M. G. and Ali, M., A Technique of Fuzzy C-Mean in Multiple Linear Regression Model toward Paddy Yield; *Journal of Physics: Conference Series*; **995(1):** Apr 2018; PP: 012010 (2018).