IJSEAT International Journal of Science Engineering and Advance Technology

A Study on Goodness of Fit For The Linear Model of Index Numbers of Jute Productivity In India

S. V. Subrahmanyam[1*], Dhaneshwar Prasad[2], S.S. Panda[3]

[1]Department of Mathematics, K. L. University, Guntur-522502, Andhra Pradesh, India,
[2]Department of Mathematics, Dr. S. R. K. Govt. Arts College, Yanam-533464, Puducherry, India.
[3]Associate Professor, Regency Institute of Technology, Yanam-533464, Puducherry, India.
*manyam19@gmail.com

ABSTRACT:
A linear trend is fitted to the index numbers of Jute productivity in India from 1993-94 to 2011-12. The line is tested for its goodness of fit by studying standard error, closeness of fit, confidence region, and some sampling distribution techniques such as chi-square test, t-test, z-test. As a result, it is concluded that there is no significant difference between the observed and estimated index numbers of Jute productivity in India from 1993-94 to 2011-12 and hence the regression line is a good fit.
Key words: Mathematical Model, Closeness of fit, Correlation coefficient, Chi-square test of goodness of fit, t-test, and z-test.

1. INTRODUCTION:

Jute is the second most important vegetable fiber after cotton. Jute is used to make cloth for wrapping bales of raw cotton, and to make sacks and coarse cloth. The fibers are also woven into curtains, chair coverings, carpets, area rugs, hessian cloth, and backing for linoleum. While jute is being replaced by synthetic materials in many of these uses, some uses take advantage of jute's biodegradable nature, where synthetics would be unsuitable. Examples of such uses include containers for planting young trees, which can be planted directly with the container without disturbing the roots, and land restoration where jute cloth prevents erosion occurring while natural vegetation becomes established. Jute is regularly used in the production of clothes, particularly sweaters and underwear. Its soft fabric structure is known to be very comfortable to the person wearing it and therefore, popularity for jute clothes has increased significantly over the last decade. The fibers are used alone or blended with other types of fiber to make twine and rope. Jute butts, the coarse ends of the plants, are used to make inexpensive cloth. Conversely, very fine threads of jute can be separated out and made into imitation silk. As jute fibers are also being used to make pulp and paper, and with increasing concern over forest destruction for the wood pulp used to make most paper, the importance of jute for this purpose may increase. Jute has a long history of use in the sackings, carpets, wrapping fabrics (cotton bale), and construction fabric manufacturing industry. Traditionally jute was used in traditional textile machineries as textile fibers having cellulose (vegetable fiber content) and lignin (wood fiber content). But, the major breakthrough came when the automobile, pulp and paper, and the furniture and bedding industries started to use jute and its allied fibers with their non-woven and composite technology to manufacture nonwovens, technical textiles, and composites. Therefore, jute has changed its textile fiber outlook and steadily heading towards its newer identity.

Index numbers are today one of the most widely used statistical devices. They are used to take the pulse of the economy and they are to be used as indicators of inflationary or deflationary tendencies. The first index number was constructed by an Italian, Mr. Carli, in 1764 to compare the changes in price for the year 1750 with the price level in the year 1500 in order to study the effect discovery of America on the price level in Italy. The index numbers help in studying trends and tendencies, in formulating decisions and policies, and are used for deflation. Further, it would help in deciding about the nature of the statistics to be collected, the statistical techniques to be used and also has a determining effect on some other related problems like selection commodities, selection of base period, the average to be used and so on [1].

In statistics, the goodness of fit problems received attention recently due to their methodology and limitations. In modern experiments, the experimenter needs to estimate how accurately the fit function approximates the observed distribution [1]. There is long standing controversy about the connection between hypothesis testing and goodness of fit problem. It can be argued that there can be no alternative hypothesis for the goodness of fit test [2]. In this approach, the experimenter does not have any criteria for choosing a goodness of fit procedure over another.

The study of the test of significance plays a vital role in sampling theory which enables us to decide on the basis of sample results if the deviation between the observed sample statistic and the hypothetical parameter value differ much. Since, for large value of the sample size n, almost all the distributions can be approximated very closely by normal probability curve, we use normal test of significance for large samples. Some of the well known tests of significance for studying such difference for small samples are t-test, z-test (Fisher's z-transformations), etc. [3]. Some of the statistical methods such as closeness of fit, correlation coefficient, confidence region, etc., have been used to test for the goodness of fit for the current data and concluded that it is a good fit.

In this paper, some statistical techniques like closeness of fit, standard error, correlation coefficient are applied. Also some sampling distribution methods, such as chi-square test, t-test, z-test etc., are used to the data which are collected from Statistical Abstract of India (Freely downloaded from web) to test the goodness of fit for the linear model of index numbers of jute productivity in India from 1993-94 to 2011-12. The collected data is furnished in Table.1.

Table 1
Index number of Jute productivity in India
(Base: Triennium ending 1993-94=100)

Index numbers of Jute productivity			
Year	**Observed**	**Year**	**Observed**
1993-94	157.2	2003-04	218.9
1994-95	170.8	2004-05	200.7
1995-96	163.9	2005-06	212.9
1996-97	212.8	2006-07	220.3
1997-98	212.7	2007-08	218.3
1998-99	188.7	2008-09	205.8
1999-00	201.3	2009-10	239.8
2000-01	199	2010-11	231.8
2001-02	226	2011-12	232.6
2002-03	219.4	--	--

2. GRAPHICAL REPRESENTATION OF INDEX NUMBERS OF JUTE PRODUCTIVITY IN INDIA:

Graphical representation provides us to have a clear understanding of the data collected. There are several graphical representations are made available, out of which we selected bar diagrams method to represent the collected data.

2.1 Bar diagram:

A bar diagram is a chart with rectangular bars whose lengths proportional to the values that they represent. A bar is a thick line whose width is shown merely for attention and can be plotted vertically or horizontally. A bar diagram displays data visually and allows viewers to compare items displayed. Data displayed will relate to things like amounts, characteristics, times and frequency etc. A bar graph displays information in a way that helps us to make generalizations and conclusions quickly and easily. Multiple bar diagram is preferred to make comparison between two or more related variables.

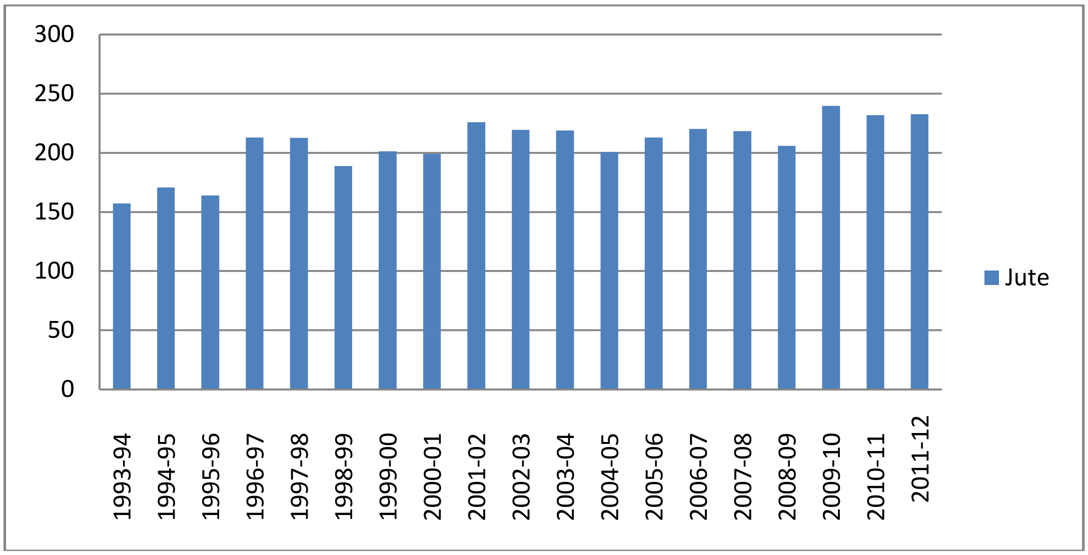

Fig. (1) : Bar diagram for Index numbers of Jute productivity in India

3. STRAIGHT LINE:

The regression line fitted to the observed values of index numbers of Jute productivity in India from 1993-94 to 2011-12 by the method of least squares is $Y = 3.23X + 174.69$; the observed and estimated values are presented in Table (2) and they are depicted in Fig. (3).

Table 2
Observed and estimated values of index numbers of Jute productivity

Index numbers of Jute productivity					
Year	**Observed**	**Estimated**	**Year**	**Observed**	**Estimated**
1993-94	157.2	177.92	2003-04	218.9	210.22
1994-95	170.8	181.15	2004-05	200.7	213.45
1995-96	163.9	184.38	2005-06	212.9	216.68
1996-97	212.8	187.61	2006-07	220.3	219.91
1997-98	212.7	190.84	2007-08	218.3	223.14

1998-99	188.7	194.07	2008-09	205.8	226.37
1999-00	201.3	197.3	2009-10	239.8	229.6
2000-01	199	200.53	2010-11	231.8	232.83
2001-02	226	203.76	2011-12	232.6	236.06
2002-03	219.4	206.99	--	--	--

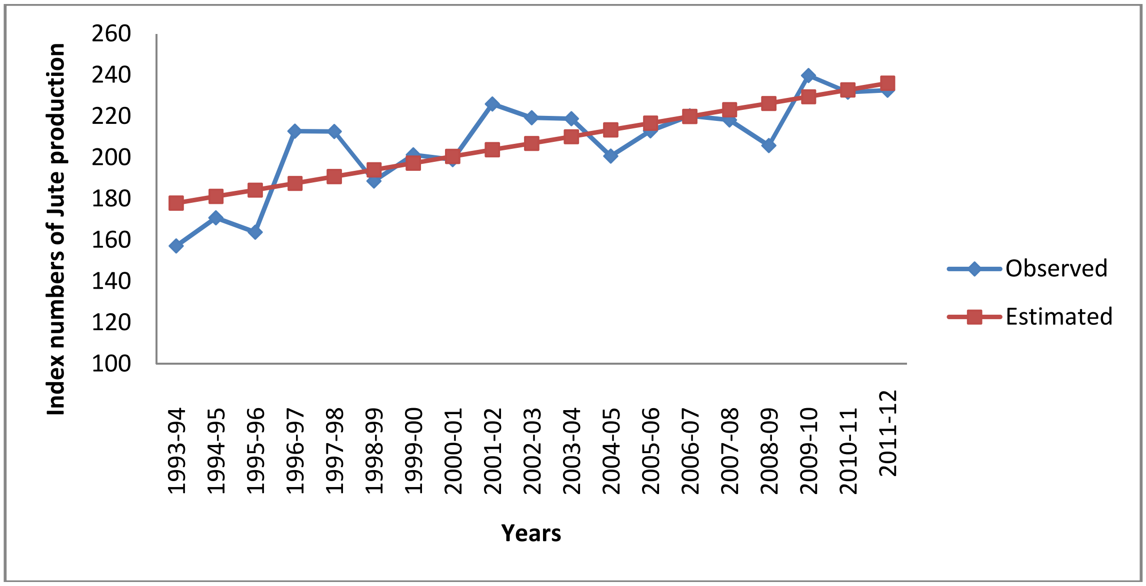

Fig. (2) : Observed and estimated values of index numbers of Jute production

4. STANDARD ERROR:

The standard error [4] for the index numbers of jute productivity can be calculated by using the formula

$$\sigma_Y = \sqrt{\frac{\Sigma (Y - \overline{Y})^2}{N-1}}$$

where N is no. of observations. Therefore, $\sigma_Y = 23$.

5. CLOSENESS OF FIT:

It is observed that how far the estimated values and the observed values are closely related. The measurement of closeness of fit [4] is studied by the formula

$$R = \sqrt{1 - \frac{U}{N\,\sigma_Y{}^2}}$$

where U = sum of the squares of the residuals = $\sum_{i=1}^{n} (O_i - E_i)^2$. Hence, R = 0.8023, and thereby the observed and estimated values are closely related.

6. CONFIDENCE REGION:

In as much as approximately 68% of the points in the scatter diagram will lie within the range of the regression line of index numbers of the jute productivity in India plus and minus one standard error provided the points scatter about this line approximately a normal distribution. A normal variable is subject to observational errors of a random sort and or the factors omitted from the function are actually numerous and uncorrelated [4]. Under these circumstances, points are located one standard error above and below the regression line of index numbers of jute productivity in India. The equations of parallel lines are worked out as $Y = 3.23X + 174.69 \pm 23$, i.e. Upper confidence line $Y_U = 3.23X + 197.69$ and Lower confidence line $Y_L = 3.23X + 151.69$, these lines are shown in Fig. (4).

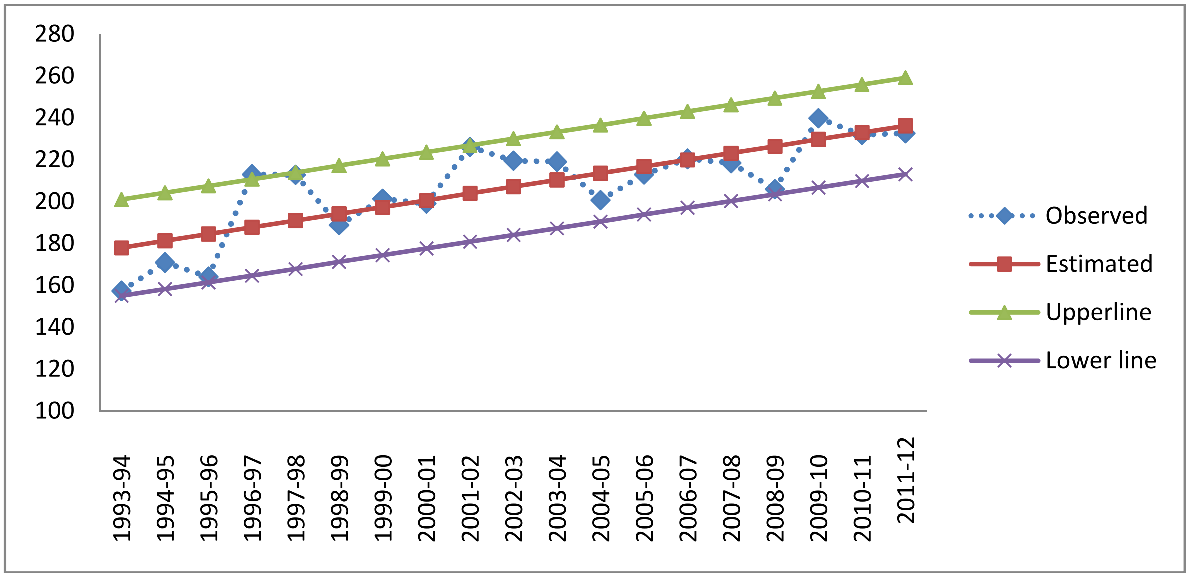

Fig.(3): Confidence region

7. χ^2 - TEST FOR GOODNESS OF FIT:

As chi-square test is the most powerful statistical tool (a non-parametric test), it is used to test the significant difference between the hypothetical and observed value of a phenomena under investigation. Chi-square analysis, in general, is used to test the similarities between two or more populations or variables on some characteristics of interest. (Other statistical test does pair wise comparison, but the chi-square can handle more than one variable or population at the same time).

The chi-square analysis is appropriate when we need to do the following:

(1) Compare the independence of two characteristics or how they are related to or independent of each other (Test of Independence)

(2) Test whether different populations are similar (or homogeneous) to some common characteristics. (Test of Homogeneity)

(3) Use to make inference about the population variance.

(4) Goodness of fit for a given data.

(5) Compare a distribution with a reference distribution such as the normal distribution.

(6) Test whether two or more distributions are identical.

(7) Compare the proportions or frequencies of categorical data (Goodness-of-Fit)

Goodness-of-fit test is a chi-square test technique used to study similarities between proportions or frequencies between groupings (or classification) of categorical data.

Now, it is assumed under null hypothesis that there is no significant difference between observed values and estimated values of index numbers of jute productivity; and alternate hypothesis is that there is a significant difference between observed and estimated values under 95% level of significance. The test statistic for χ^2 - test [5] for testing our null hypothesis is

$$\chi^2_{(n-1,0.05)} = \sum_{i=1}^{n} \frac{(O_i - E_i)^2}{E_i}$$

where n is number of values, O_i and E_i are observed values estimated values of index number of jute productivity respectively.

Thus, one can obtain the calculated χ^2 value as 18.2474, but the tabulated χ^2 value for 18 degrees of freedom at 95% level of significance is 28.869. Since calculated χ^2 value is less than tabulated χ^2 value, one can accept the above null hypothesis. Thus by χ^2 test, one can conclude that the above fitted straight line is a good fit.

8. CORRELATION COEFFICIENT (r):

As a measure of intensity and degree of linear relationship between two or more variables of a given data, correlation is the most powerful statistical tool given by a famous Bio-Metrician Prof. Karl Pearson. According to Karl Pearson, correlation coefficient is a number which lies between -1 and 1 which measures the degree to which two variables are linearly related. If there is perfect linear relationship with positive slope between the two variables, we have a correlation coefficient of 1; if there is positive correlation, whenever one variable has a high (low) value, so does the other. If there is a perfect linear relationship with negative slope between the two variables, we have a correlation coefficient of -1; if there is negative correlation, whenever one variable has a high (low) value; the other has a low (high) value. A correlation coefficient of zero means that there is no linear relationship between the variables. The correlation coefficient (r) [5] is defined by the formula

$$r = \frac{Cov\ (Y_0, Y_e)}{\sigma_{Y_0}\ \sigma_{Y_e}}$$

where Y_0 and Y_e denote the observed and the estimated index number of jute productivity. By using the above formula, the correlation coefficient between actual values and estimated values obtained by using regression line is r = 0.79, thereby it can be inferred that the observed and estimated index numbers of jute productivity are correlated.

9. TEST FOR SIGNIFICANCE OF CORRELATION:

The test of significance of correlation coefficient is as similar as that of the test of significance of means. One difference is that the sampling distribution of r is more complex than sampling distributions for means or mean differences. The shape of the sampling distribution of r tends slowly towards the normal distribution as N increases. Significance tests allow analysts to assess whether apparent relationships between random variables are real or due to chance. In this paper we tested the significance of the relationship between the observed and estimated index numbers of jute productivity in India by using the t-test and z-test. We fix the null hypothesis that there is no correlation between the observed and the estimated index numbers of jute productivity by assuming that both the variables are normally distributed.

9.1 t - TEST FOR COEFFICIENT OF CORRELATION:

A t-test can be used to test whether the correlation between two variables is significant. The test statistic follows a t-distribution with n − 2 degrees of freedom. The test statistic for t-test [5] is

$$t_{(n-2,0.05)} = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}}$$

where n is number of values and r is correlation coefficient between O_i and E_i. Thus one can obtain the t-value as 5.3126, but the tabulated t-value for 17 degrees of freedom at 95% level of significance is 2.11. Hence, the null hypothesis is rejected as calculated t-value is greater than tabulated t- value.

9.2 Z- TEST FOR COEFFICIENT OF CORRELATION:

A z- test can be used to test whether the correlation between two variables is significant. The test statistic follows a normal distribution. The test statistic for z – test [5] is

$$z_{0.05} = \frac{\sqrt{n-3}}{2} \ln\left(\frac{1+r}{1-r}\right)$$

Thus one can obtain the z value as 4.2857, which is large when compared with the significant value (1.96) at 95% level of significance. Thus, one can reject the above null hypothesis.

From the above t-test and z- test, the null hypothesis for the index numbers of jute productivity in India is rejected. Hence by either test (t- test, and z-test) one can conclude that there is no significant difference between the observed and the estimated index numbers of jute productivity and hence the above fitted line of regression is a good fit.

10. CONCLUSIONS:

A linear regression $Y = 3.23\,X + 174.69$ is fitted for the index numbers of jute productivity from 1993-94 to 2011-2012 in India and tested for its significance of goodness of fit by its correlation between the observed and estimated values (obtained from the fitted straight line), closeness of fit, confidence region, χ^2 - test, t-test, z - test. It is concluded that there is no significant difference in between the observed and estimated values of index numbers of jute productivity in India and hence the fitted line of regression is a good fit.

11. REFERENCES:

[1]. S.C. Gupta, 2005, "Fundamentals of Statistics", 6th edn, Himalaya Publishing House, India

[2]. I.Narsky, 2003,"Goodness of fit: what do we really want to know?", PHYSTAT2003, SLAC, Stanford, California, September 8-11,

[3]. S.C.Gupta & V.K.Kapoor, "Fundamentals of mathematical statics", S.Chand & Co., New Delhi.

[4]. K. Venkateswara rao & S.V. Subrahmanyam, 2012, "A mathematical model of rice productivity in Andhra Pradesh", International Journal of Mathematical Archive, Vol.3(11), PP. 4047-4051.

[5]. Richard A. Johnson, 2007, "Miller & Freund's Probability and Statistics for Engineers", New Delhi, Prentice – Hall of India (P) Ltd., 7th ed.