A STUDY ON CAPARISON OF SOME STATISTICAL MODELS

S. Loidang Devi
Department of Statistics
D.M. College of Science, Imphal, Manipur, India.

Abstract: This paper consists of two models. One is usual multiple linear regression model and the another is principal component regression model. Primary data is used for this analysis. Findings show that the models fit the data well. Diagnostic checks confirmed that data do not seem to contradict the general underlying assumptions about the models. The value of Multiple correlation found in the usual regression model suggest that out of the total variation, 92.4% of variation in the yield of paddy is explained by the independent variable under consideration whereas 88.1% of variation is explained by principal component under study as shown by the principal component regression model.

Key words and phrases: Statistical Model, Stratified two-stage sampling scheme, Primary data, Multicollinearity, etc.

1. Introduction

The yield of crop depends on multiple of factors acting collectively. The collective influence of these factors may not be systematically arranged and properly managed (by the farmers) as because one factor may increase while some other may decrease and still another remain constant. Specially, the behavioral nature of these factors for their contribution to the yield is not predictable in general. In view of this consideration, we are trying to find the factors that have signif icant influences on the yield of rice so that we can control it for increasing production if it is controllable. A number of workers (Parikh and Mosley,1986; Mohapatra et al, 1996; Rajinder Kaur and Sekarwar, 1997) investigated the influence of different factors on the production of crop in the past. The existing models need to be modified to suit the prevailing conditions of this region as well as to enable us to carry out formal mathematical analysis using relevant and sophisticated data. On this basis, we have been trying to build a statistical model that would enable us to identify the important factors influencing rice production and thus controlling it for better production in future.

2. The Model

The statistical model proposed for the production of rice is

$$\log_e Y_i = b_{io} + \sum_{j=i}^{p} b_{ij} \log_e X_{ij} + \varepsilon_i, \quad i = 1,2, \ldots, n; \; j = 1,2, \ldots, p \text{ ---------- (1)}$$

where, Y_i = Yield of the i^{th} farm in kilograms,
X_{i1} = Age of the i^{th} farmer in years,
X_{i2} = Year of schooling of the i^{th} farmer,
X_{i3} = Size of the family for the i^{th} farmer,
X_{i4} = Area under tenant operated by the i^{th} farmer in hectare,
X_{i5} = Area under owner operated by the i^{th} farmer in hectare,
X_{i6} = Area under double cropping of the i^{th} farmer.
X_{i7} = Irrigated area of the i^{th} farmer (if any) in hectare,
X_{i8} = Quantity of fertilizers consumed by the i^{th} farmer in kilograms,
X_{i9} = Area under modern High-Yielding Variety in the i^{th} farm in hectare,
X_{i10} = 1, if the soil of the i^{th} farmer has been tested,
= 0, otherwise,
X_{i11} = Quantity of farmyard manure applied to the i^{th} farm (per load of bull card),
X_{i12} = Distance of market from the i^{th} farmer's residence,
X_{i13} = Amount of loan availed by the i^{th} farmer from any govt. recognized agency, etc.,
X_{i14} = Cost of cultivation excluding fertilizer cost incurred by the i^{th} farmer,
X_{i15} = Monthly family income of the i^{th} farmer in rupees,
b_{i0} = constant,
b_{ij} = regression co-efficients,
ε_i's are error components and are assumed to be independently and identically distribute as $N(0, \sigma^2)$.

(All the variables are involved log except the variable X_{i10}).

3. The Sampling frame

To fit the model, we collect the cross-sectional data by preparing a pre-designated questionnaire method is used. The sampling design of this crop survey is a stratified two stage sampling scheme of equal size suggested by Cochran (1977). With the blocks of Imphal West District as strata, villages in the blocks as the primary sampling unit and farmers of experimental site of the selected villages as the ultimate second stage sampling unit. The dated used in the study were from the survey of 795 farms in 43 villages.

4. Data Analysis

In order to analysis the data, we used the SPSS-software packages. Ordinary least squares method is used for fitting the model.

Results

(a). For multiple linear regression model:

TABLE – 1: Estimated Regression Parameters, t-statistic, multicollinearity Statistics

Variables	Coefficients(b)	t	Collinearity Statistics	
			Tolerance	VIF
(Constant)	-1.682	-11.120		
X_{i1}	2.983E-02	.981	.811	1.233
X_{i2}	1.372E-03	.180	.775	1.290
X_{i3}	4.001E-02	1.659	.886	1.129
X_{i4}	.279	12.967*	.311	3.218
X_{i5}	.287	13.743*	.282	3.541
X_{i6}	9.929E-02	2.550*	.854	1.172
X_{i7}	5.057E-03	.447	.855	1.169
X_{i8}	.152	11.180*	.412	2.427
X_{i9}	.154	7.903*	.472	2.117
X_{i10}	.155	2.152*	.979	1.021
X_{i11}	3.670E-02	4.449*	.906	1.104
X_{i12}	1.282E-02	1.245	.881	1.135
X_{i13}	5.186E-03	.742	.992	1.009
X_{i14}	8.853E-02	7.375*	.517	1.935
X_{i15}	6.908E-02	8.428*	.763	1.311

***** indicates significant at 5% level.

Analysis of data shows that majority of the regression co-efficients, (X_{i4}, X_{i5}, X_{i6}, X_{i8}, X_{i9}, X_{i10}, X_{i11}, X_{i14}, X_{i15}) are significant while others are found insignificant. So, it confirms that the linearity condition between Y_i and X_{ij}'s, $\forall$ j, may reasonably be assumed in the model (Table-1). Further, it can be observed that the tolerance limits are not small, and each is not less than 0.01 and the VIF(Variance Inflation Factor) is not large, and each is not greater than 100, and these results together imply that multicollinearity is absent.

TABLE –2 ANOVA

Sources of variables	Sum of Squares	df	Mean Square	F
Regression	137.427	15	9.162	239.351**
Residual	29.819	779	3.828E-02	
Total	167.247	794		

****** indicates highly significant at 5% and 1% level

Since the calculated F-value namely 239.351 is highly significant, we reject the null hypothesis that the b_{ij}'s are zero at the 0.05 and 0.01 level of significance even. Hence, each Y_i can be predicted by the X_{ij}'s accurately.

TABLE – 3: Multiple Correlation co-efficient (R), F-statistic and Durbin-Watson (d) statistic

R	R- Squared	Adjusted R- Squared	F- value	df1	df2	Durbin-Watson statistic
.906	.822	.818	239.351	15	779	1.681

It follows from the above Table-3, the computed value of the Durbin-Watson statistic (which is 1.681) is very close to 2, and this indicates that the residuals ε_i's are independent.

The scatter plot, Fig.3 is of the type of ellipse of concentration, and it confirms the absence of heteroscedasticity of the variances of the residuals.

Since the normal probability plot of standardized residuals, shown in Fig.2 is close to the diagonal of the box, it may be asserted that the residuals follow a joint multivariate normal distribution and hence it may be concluded that the data can form a normal population.

Furthermore, it may be observed in Fig.1 that almost all the position of histogram is superimposed by normal density curve, and this clearly indicates that the data are drawn from a normal distribution.

Q-Q plot for each variable also confirms that all the variables follow a normal distribution.

After removing the outliers and retaining only significant variables, we have run the analysis as usual.

TABLE – 4: Estimated Regression Parameters, t-statistic, multicollinearity Statistics

Variables	Coefficients(b)	t	Collinearity Statistics	
			Tolerance	VIF
(Constant)	-1.240	-12.690		
X_{i4}	.392	18.251*	.261	3.828
X_{i5}	.413	19.153*	.229	4.370
X_{i6}	.139	4.160*	.916	1.091
X_{i8}	.125	10.133*	. 261	2.525
X_{i9}	.113	6.382*	.463	2.161
X_{i10}	.172	2.734*	.992	1.009
X_{i11}	3.125E-02	4.318*	.934	1.071
X_{i14}	7.673E-02	7.193*	.522	1.915
X_{i15}	5.920E-02	8.340*	.813	1.230

***** indicates significant at 5% level.

Analysis of data shows that all the regression co-efficients, (X_{i4}, X_{i5}, X_{i6}, X_{i8}, X_{i9}, X_{i10}, X_{i11}, X_{i14}, X_{i15}) are significant. So, it confirms that the linearity condition between Y_i and X_{ij}'s, $\forall$ j, may reasonably be assumed in the model (Table-5). Further, it can be observed that the tolerance limits are not small, and each is not less than 0.01 and the VIF (Variance Inflation Factor) is not large, and each is not greater than 100, and these results together imply that multicollinearity is absent.

TABLE – 5

Multiple Correlation co-efficient (R), F-statistic and Durbin-Watson (d) statistic

R	R- Squared	Adjusted R- Squared	F- value	df1	df2	Durbin-Watson statistic
.924	.824	.853	506.472	9	776	1.713

It follows from the above Table-6, the computed value of the Durbin-Watson statistic (which is 1.713) is very close to 2, and this indicates that the residuals ε_i's are independent.

(b). For principal component regression model:

First we extract fifteen (15) principal components. Out of these, only six components are selected as the variance extracted being one and greater [by the criterion of 'root greater than one' by Kaiser (1958)]. By using these six principal components, the model in equation (1) is fitted by ordinary least square method.

TABLE-1: Estimated Regression Parameters, t-statistic, multicollinearity Statistics

Variables	Coefficients (b)	T	Collinearity Statistics	
			Tolerance	**VIF**
(Constant)	-2.372	-25.656		
Y_{i1}	0.184	45.356*	0.810	1.234
Y_{i2}	8.779E-05	3.756*	0.030	33.712
Y_{i3}	-2.268E-02	-2.218*	0.744	1.345
Y_{i4}	2.211E-02	1.742	0.774	1.292
Y_{i5}	-7.423E-02	-4.449*	0.480	2.081
Y_{i6}	0.215	3.956*	0.030	33.480

* indicates significant at 5% level.

Analysis of data shows that majority of the regression co-efficients, (Y_{i1}, Y_{i2}, Y_{i3}, Y_{i4}, Y_{i5}, Y_{i6}) are significant except Y_{i4}. So, it confirms that the linearity condition between Z_i and Y_{ij}'s, may reasonably be assumed in the model (Table-1). Further, it can be observed that as tolerance limits are not small, and each is not less than 0.01 and the VIF(Variance Inflation Factor) is not large, and each is not greater than 100, and these results together imply that multicollinearity is absent.

TABLE-2: ANOVA

Sources of variables	Sum of Squares	df	Mean Square	F
Regression	130.908	6	21.818	453.789**
Residual	37.887	788	0.48	
Total	168.795	794		

** indicates highly significant at 5% and 1% level.

Since the calculated F-value namely 453.789 is highly significant, we reject the null hypothesis that the b_{ij}'s are zero at the 0.05 and 0.01 level of significance even. Hence, each Z_i can be predicted by the Y_{ij}'s accurately.

TABLE-3: Multiple Correlation coefficient ®, F-statistic and Dorbin-Watson (d) statistic

R	R-Squared	Adjusted R-Squared	F-value	df1	df2	Durbin-Watson statistic
0.881	0.776	0.774	453.789	6	788	1.590

It follows from the above Table-3, the computed value of the Durbin-Watson statistic (which is 1.590) is very close to 2, and this indicates that the residuals $\mathcal{E}_i$'s are independent.

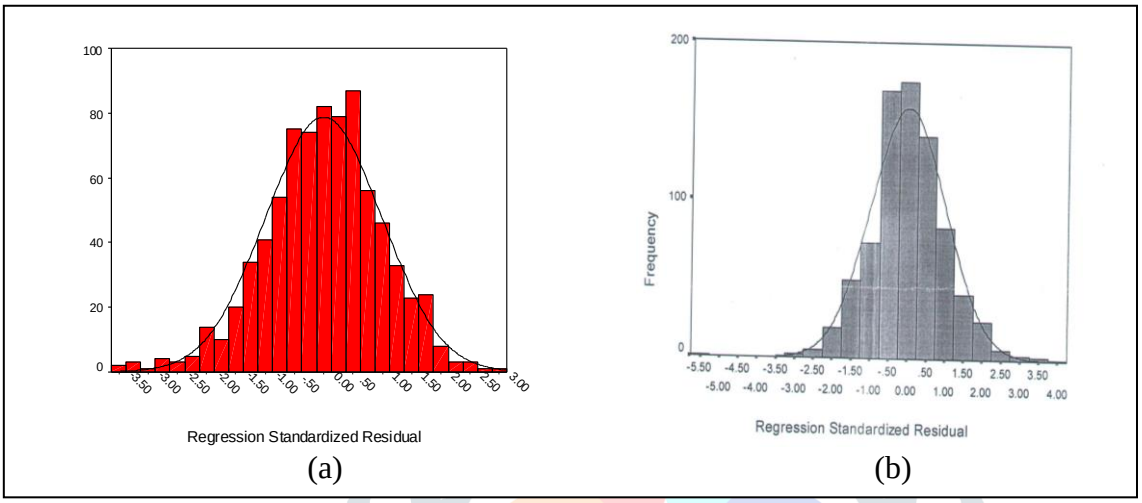

Fig.1-Histogram and Normal Density Curve for Normality Test of Sample Data

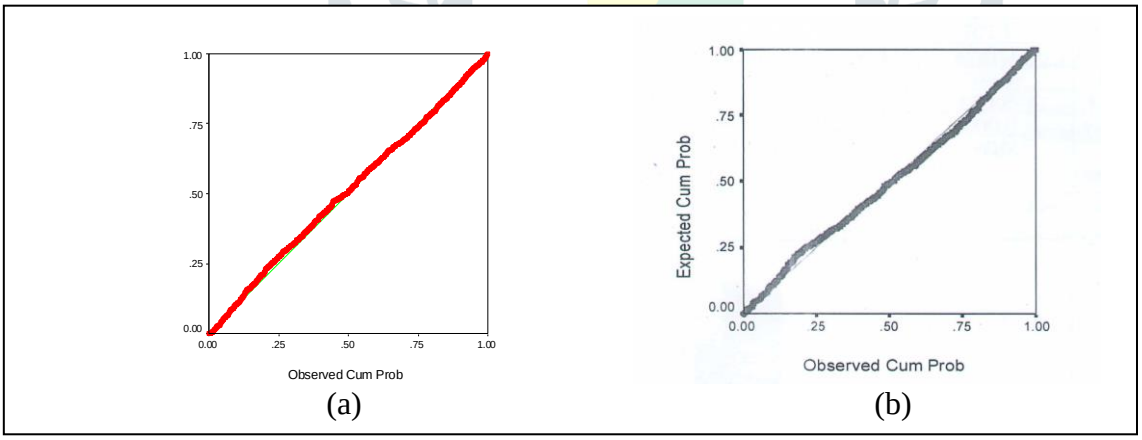

Fig.2-Normal Probability Plot of Regression Standardized Residual

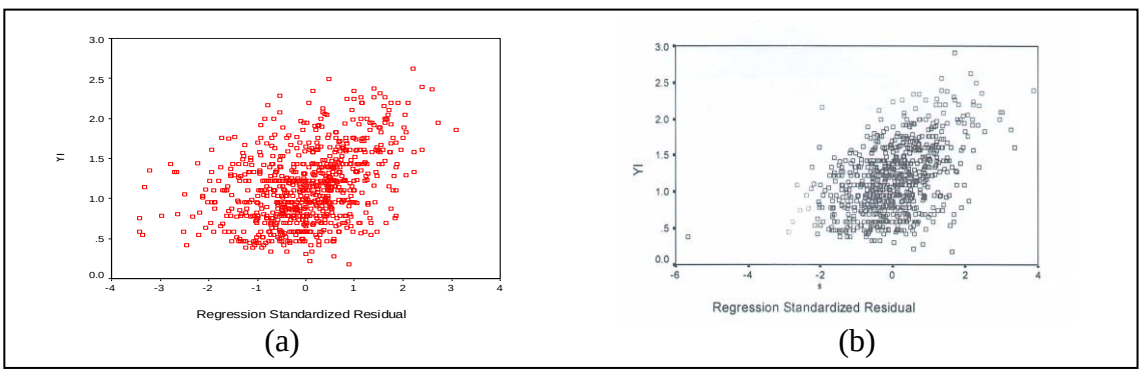

Fig.-3: Homoscedasticity of Residual

5. Conclusions

The proposed model was found to satisfy all the assumptions and criteria, Further, testing of the above hypotheses and diagnostic checks confirm that the models fit the data well.

The considered independent variables could predicts the values of the dependent variables. For the first model, it was R = 0.924 and for the second model, it was R = 0.881. On comparing these two models, the first model is more efficient.

To be precise, the following had been either experienced or observed:

(1) Fertilizers, as it is generally expected, have a significant effect on the yield. This finding along with the farmer's wailings for inadequate amount of fertilizers serves a pointer to the fact that, even though we did achieve substantial progress through special programmes for developing small farm agriculture, there exists further scope to increase utilization of fertilizers. So, it demands for opening more fertilizer cells and sale branches in the state.

(2) Cropping pattern also substantially influences the increase of production. This suggests that more and more portion of cultivable area must be allowed to enter under this pattern, if possible so that the production rapidly increases.

(3) This study also suggests that HYV's be sown on a larger scale and more and more area be brought under this to boost crop yield. For this, infrastructural facilities are indispensable in Manipur.

This is concerned with crop estimation practices (for planning and policy making). The findings point out that the models may be used as crop production model in the region.

Lastly, it is sincerely felt that such studies are essentially called for such underdeveloped regions like Manipur and that with the increase in the availability of scientific and technological facilities there will come up a wider and wider scope of modification and application of the models and the statistical tools and techniques used in the proposed study.

References

1. Afifi, A.A. and Clark, V. (1984): *Computer –Aided Multivariate Analysis*. Chapman & Hall, London.
2. B.M. Singh (2002): Multivariate Statistical Analysis; An introduction to its theoretical Aspects. South Asian Publisher PVT Ltd. New Delhi.
3. Census of India (1991): District Census Handbook, Imphal, Director of Census Operations, Manipur.
4. Cochran, W.G (1977): *Sampling Techniques*. John Wiley, New York.
5. Dholakia, R.H. and Majumdar, J. (1995): Estimation of Price Elasticity of Fertilizer Demand at Macro Level in India. *Indian Jn. of Agri. Econ*. (Jan.-March, 1995), Indian Society of Agricultural Economics, Bombay, Vol. 50, No.1. Page 36-45.
6. Economic Review (1994-95): Directorate of Economics and Statistics, Govt. of Manipur., Imphal.
7. Kaur, R. and Sekarwar, H.S. (1997): A Statistical Study to Evaluate the Relative Performance of Different Sources of Phosphorus in a Rice-Rice Cropping System. Annals of Agricultural Research (September, 1997), Indian Society of Agricultural Science, New Delhi, Vol. 18, No. 3, pp. 285-289.
8. Parikh, A. and Mosley, S. (1986): Fertilizer Response in Haryana. *Ind. Jn. of Agri. Econ.* (April-June, 1986), Indian Society of Agricultural Economics, Bombay, Vol. 41 No. 2, pp. 141-153.
9. *Report on Agricultural Census, (1980-81) and Input Survey, (1981-82) of Manipur*: Department of Agriculture, Govt. of Manipur, Imphal.
10. Statistical Abstract of Manipur, 2001: Directorate of Economics and Statistics, Govt. of Manipur.
11. Younger, M.S. (1979): *A Handbook of Linear Regression*. Wadsworth, Inc., Belmont, California.