UNIVERSITY OF ZAMBIA
DEPARTMENT OF MATHEMATICS
LUSAKA, ZAMBIA

STATISTICAL ANALYSIS OF DAILY RAINFALL DATA

BY

NCHIMUNYA URBAN HAANKUKU

DISSERTATION FOR
M.Sc. (MATHEMATICS)
×WITH STATISTICS×

NOVEMBER 1989

DECLARATION

This dissertation represents my own work and it has not previously been submitted for a degree at this or any other University.

Signature: ....

Date: ....8-10-90...........................

APPROVAL

This Dissertation of Nchimunya Urban Haankuku is approved as fulfilling part of the requirements for the award of the degree of Master of Sciences in Mathematics with Statistics.

1. Signed :________________________ Date : 26th July 1989
(EXTERNAL EXAMINER)

2. Signed :________________________ Date : 18th Sept. 1989

3. Signed :________________________ Date :________________

4. Signed :________________________ Date :________________

ACKNOWLEDGEMENTS

I AM DEEPLY INDEBTED AND WOULD LIKE TO EXPRESS MY HEARTFELT GRATITUDE TO MY SUPERVISOR PROFESSOR G. EVANS (PROFESSOR OF STATISTICS, STATISTICS DEPARTMENT, WALES UNIVERSITY - LONDON) FOR INITIATING THIS PROJECT, AND MY INTERNAL SUPERVISOR DR. J. PRAKASH (SENIOR LECTURER, DEPARTMENT OF MATHEMATICS - UNIVERSITY OF ZAMBIA, LUSAKA) WHO HAVE CONSTANTLY ENCOURAGED AND GUIDED ME THROUGHOUT THE PREPARATION OF THIS DISSERTATION. THEIR CONSTRUCTIVE CRITICISMS AND VALUABLE SUGGESTIONS HAD BEEN A GREAT SOURCE OF INSPIRATION THROUGHOUT THIS WORK.

I EXPRESS MY SINCERE THANKS TO ALL MEMBERS OF THE DEPARTMENT OF MATHEMATICS (UNIVERSITY OF ZAMBIA) IN PARTICULAR PROFESSOR M. DUMMER, DR. KALENGA FOR THE ASSISTANCE THEY GAVE ME IN MY ADDITIONAL CHAPTER ON "ANALYSIS OF RAW DATA".

I AM ALSO GRATEFUL TO MR. N. NG'ANDU (LECTURER, DEPARTMENT OF COMMUNITY HEALTH, SCHOOL OF MEDICINE - UNIVERSITY OF ZAMBIA, LUSAKA) FOR HIS TIMELY ASSISTANCE HE GAVE ME ON THE USE OF THE COMPUTER.

MY THANKS GO TO MISS S. P. MOONO FOR TYPING THE MANUSCRIPT AND LASTLY MY THANKS GO TO MR. T. MULELE FOR GRAPHICS ASSISTANCE HE GAVE ME.

NOVEMBER 1989 **NCHIMUNYA URBAN HAANKUKU**

I would like to express my heartfelt gratitude to my wife Victoria Kabunda for her courage of looking after my family during my absence from home, her patience for a long period I was away from home. Also I would like to express my gratitude to my children, Busiku, Mazuba, Chilala and Ceasor who tolerated my absence from homefor a long period while pursuing my studies.

I would also like to express my gratitude to my brother Joseph (late), Richard, Bartholomeusz and Everisto for the encouragement they kept on giving me during my times of difficulties during my studies. I would also like to express gratitude to a friend of mine A. Ngwengwe in America who kept on helping me in study materials and encouragement dispite the distance. Also I would like to express gratitude to Mr. N. Ng'andu for his assistance on computer assistance and encouragement he gave to me.

CONTENT

LIST OF SYMBOLS

1.	X_1'	=	The smallest sample value
2.	δ	=	Random error
3.	SSE	=	The residual sum of squares
4.	MS_E	=	The residual mean square
5.	S_{yy}	=	The corrected sum of squares of y
6.	β	=	A (PX1) vector of the regression coefficients
7.	α	=	Level of significance
8	Y_o	=	Future observation when $X = X_o$
9.	$\hat{Y}_o$	=	The predicted future observation when $X = X_o$
10	X_o	=	The vector of the level of regressor variables
11	X	=	(nxP) matrix of the levels of the regressor variables
12	Y	=	(nxP) vector of the dependant variable

CHAPTER I

GENERAL INTRODUCTION

1.1 INTRODUCTION

Throughout the world considerable effort is deveoted to the collection of rainfall data. Many long records exist and most countries including Zambia, now have a reasonably dense network systems of rainfall stations. The work on data collection is not at present matched by a corresponding effort on analysis. This project demonstrates methods of analysis of daily rainfall records and shows how the models derived can be used to give results that are of direct use in Agricultural planning and designing. In many parts of the world, the climate is the major constraint on agriculture and agricultural planning decisions must be made with this in mind. Although various climatic variables interact with the crop in complex ways, RAINFALL is the crucial factor in most parts of the tropics.

The basic intention of this project is to make predictions about the future rainfall using appropriate statistical procedures with the available rainfall data at a fixed Weather Station. The Agricultural and designer's important features of the rainfall patterns pertaining to a given year include the following

(a) X_1 = The start of the rain season (meteoro-Definition) measured in days from September 23rd.

(b) X_2 = The first day of at least two consecutive days with rainfall and the accumulated is 10mm or more measured in days from September 23rd

(c) X_3 = The total rainfall from September 23rd to November 30th.

(d) X_4 = The total rainfall from September 23rd to December 31st.

(e) Y_1 = The end of the rainy season (meteorological definition) in days from September 23rd.

(f) Y_2 = The total rainfall for the season.

METEOROLOGIST, define the beginning/end of the season as:

Beginning of season: A block of seven days with the first three days with rainfall and the total sum of rainfall in this block equal or exceed 20mm. The first day is the beginning of the rain season.

End of Season: A block of seven days with the last three days with rainfall and the total sum of rainfall in this block equal or greater than 20mm. The last day marks the end of the rainy season.

In most cases, the beginning/end of the rainy season is user dependant.

Given the available data, it is necessary to begin by testing whether the given data display a time trend. If a significant time trend is observed, then a time series approach is suggested for the analysis of the data. But if such a time trend is not evident, a classical approach can be used which involves the construction of models that reflects the available data as closely as possible.

The run tests for randomness and serial correlation indicated that the data used in this project may be assumed to be purely random. Therefore, the classical approach was used for the analysis.

There is no particular reason why any one particular distribution should be used since there has been no study made of the various possibilities. All we can say is that distribution with two parameters is preferable for greater flexibility. Possible distribution to use for this purpose include

(a) Lognormal

(b) Gamma

(c) Weibull

(d) Truncated Exponential

For analytic simplicity in this project, we have chosen to use the Truncated expontential distribution whose density is given in Section 3.4. The advantage of this choice is that Lawless J. F. (9) has derived expression for upper 100p per cent bounds for a predicted value. Unfortunately, the predictions obtained were not particularly informative due to the large variations present. It is possible that one of the other possible distribution would be more fruitful but time did not permit this to be investigated. The above approach is aimed at predictions about a future rainy season. For within-a-season predictions, linear regression (both simple and multiple) were used.

In this project, predictions of Y were investigated using regression analysis for the given independent X variables. These models, that is, the truncated exponential distribution and linear/multiple regression analysis were tested for model adequacy before they were used. At this time, transformation of the variable might be considered. The measure of model adequacy for regression analysis given in section 4.3 help to detect outliers and to give ideas on whether transformations of variables were necessary. In this project, section 4.3 shows that the models used for regression analysis were adequate. Hence, there was no need of transformation of variables.

The predictions of the dependent variables using linear regression analysis are given in section 4.4 and 4.5. Sections 4.5.2.1 -5 shows the results of the predicted dependent variables using multiple regression analysis. The general result shows that section 4.5.1.1 and 4.5.1.2 give the same predictions. For linear regression analysis, it was observed that the predicted variables were basically the same for with-in-seasons predictions for both linear and multiple regression analysis. Using the results given in MONTGOMERY D.C. and PECK E.A. (10) point and interval predictions for with-in-season for both linear and multiple regression analysis were obtained.

1.2 SURVEY OF PREVIOUS WORK

Some work of restricted scope has been done on the analysis of rainfall in Zambia by some research workers. In many areas of the seasonally arid tropics, crops must be planted early and the date of the start of the growing season may coincide with the first heavy rainfall. DAS J. C. (3) observed the onset and retreat of the rainy season in Zambia using the movement of the Congo Air and Inter-tropical Convergency Zone. The agricultural important features of the rainfall pattern are the dates of the start and the end of the rainy season. If the start of the rainy season occurs late and the end of the rainy season occurs too soon, the crops may not have sufficient water to reach maturity. Meteorological Notes, series A illustrates the analysis of daily rainfall in Zambia by fitting a gamma distribution, and the Fisher Tippett type I distribution. DAS J.C. (4) observed the amount of the rainfall that fell per hour at a given station and fitted a gamma distribution.

The time distribution of the rainspell and the relationship between observed daily and twenty-four maximum rainfalls were investigated by DAS J.C. (5). BAILEY M. (1) observed the annual maximum daily rainfally recorded at thirty stations in Zambia and the expected maximum daily amounts for selected periods were estimated. To this annual maximum daily rainfall data, a Fisher Tippett type I distribution was fitted, and the adequacy of Fisher Tippett type I distribution Curve was justified by a statistical sign test, and a Wilcoxon Signed rank test.

Simango V.A. and DAS J.C. (11) used annual rainfall deciles for assessing drought in Zambia. Drought was deemed to have occured when the annual rainfall of a station was equal to or below 50 per cent of the normal rainfall. ISMAL S.A. (8) observed the abnormal amounts of rainfall during the rainy season 1972/73 for all stations in Zambia. Daily rainfall is recorded at each weather station at 08.00 hours Zambian time. A 2 day rainfall might have occured in a couple of hours only a short period before the hour of observation and another short period after the hour of observation. In these circumstances, a reference to the distribution of rainfall per hour in a day using the average method would be misleading. DOORENBOS J and PRUITT W. O. (6) illustrated other important variables other than rainfall. Unfortunately, at many sites, reliable records for other variables are available for only a few years. Hence, many simple water balance models such as COCHEME FRANQUIN P. and COCHEM J. (2). HUTCHINSON P (7) incorporated actual rainfall data in different years together with average avaporation figures.

1.3 APPLICATION OF THIS PROJECT

Zambia, where annual rainfall is moderate to meagre over large parts of the country, and year to year variations are large, unpredictable weather can prove to be a serious obstacle to stable and systematic agricultural production. For effective agricultural planning on a rational and scientific basis, appropriate analysis of rainfall procedures need to be set up. Fairly accurate predictions within a given season or future seasons are of great importance to agriculturalists and designers (such as builders of reservoirs, dams flood control works, airports). Agriculturalists and designers need to know with confidence when the rainy season is expected to begin/end and how much rain is likely to fall in a given season. They may also want to have reasonable expectations of wet and dry spells.

This project tries to provide answers to the following questions:

(a) When is the start of the rain season expected?

(b) When is the rainy season expected to end?

(c) What is the expected total rainfall for part or the whole of the rainy season?

(d) How much rain may be expected during a season given data on the early part of a rainy season?

1.4 SOURCE OF DATA

Daily rainfall data recorded at Chingombe weather station

(Kabwe) over a period of 25 years was used as the source of data to be analysed. Rainfall data from 1st September, 1960 to 30th June, 1985 were available for the project. It is evident from the data that there is little or no rainfall in the period from 30th June to 31st August in any one year. Consequently, rain at this station is restricted to the period from early September of one year to the middle of May of the next year at latest. This period, therefore, more than adequately covers the rainy season at this station.

1.5 GENERAL METHODOLOGY

This project has three distinct parts: the first part of the analysis is found in Chapter II. In this Chapter the ANALYSIS OF RAW DATA is outlined, especially the distributions of rainfall over each rainy season for the period of 25 years.

The second part of the analysis is found in Chapter III. In this chapter the predictions of future seasons were considered assuming that each variable had a truncated exponential distribution.

The third part of the analysis is found in Chapter IV where with-in-season predictions were considered using regression analysis. In this Chapter, the rainfall data for the early part of the 1985/86 season was available, and predictions as to the other aspects of the 1985/86 rainy season were attempted using tested statistical procedures. The analysis of variance tables for all the variables, the plot of standardized residuals versus predicted standardized dependent variables for measures of model adequacy were obtained using the Statistical PACKAGE for the SOCIAL SCIENCES (SPSS) Computer package.

CHAPTER II

ANALYSIS OF RAW DATA

2.1 INTRODUCTION

Fairly accurate predictions of the beginning/end of the rainy season, distribution of the total rainfall in any given rainy season would assist in planting decision, water management and transportation of agricultural produce. The period of maturity for various crops differ from one crop to another. Table 2.1a below shows various crops grown in Zambia and their maturity periods.

CROP	MATURITY PERIOD
MAIZE	
Variety	
MM501	130 - 135 days
MM502	140 - 145 days
MM504	135 - 140 days
MM601	140 - 145 days
MM603	145 - 150 days
MM604	145 - 150 days
MM752	155 - 160 days
MMV600	130 - 135 days
MMV400	100 - 110 days
POTATO	MATURITY PERIOD
Variety	
ARKA	115 - 125 days
BARAKA	125 - 130 days
PENT LAND DELL	115 - 125 days
PIMPERNELL	130 - 140 days
UP-TO-DATE	115 - 125 days

CROP	MATURITY PERIOD
SOYABEANS	
Variety	
SANTA ROSA	110 days
KALENGA	115 days
TANIA	120 days
HERMON 147	125 days
MAGOYE	130 days
JUPITA	140 days
SORGHUM	
Variety	
ZSV - 1	115 - 125 days
FRAMIDA	100 - 110 days
SUNFLOWER	
Variety	
CCA 75	130 - 135 days
CCA 81	110 - 115 days
CH 258	105 - 110 days

Maize is the major cereal staple in Zambia. It is a sub-tropical plant which prefers hot sunny conditions with reliable and evenly distributed rainfall. The crop is grown, extensively in the Southern, Central and Eastern Provinces where soils and climate are conducive to high yields. In the Northern higher rainfall areas, and in the drier parts of Western Province and the Gwembe and Luangwa Valleys, yields are lower and production is less.

For a maize grower to plan properly on when to plant, the type of variety to be used will depend on the beginning/end of the rainy season, and the distribution of the total rainfall in that rainy season. Maize is one crop which needs enough moisture in the soil at the time of germination and pollination otherwise the crop suffer from moisture stress and that would result in poor yields. At present the planting decision are often associated with beliefs about annual rainfall patterns which haven't been objectively studied. People have beliefs such as: Early beginning of the season means early ending of the season, late beginning means late ending of the season. On the bases of these beliefs, people normally make their planting decisions. In some years rainfall may have been adequate if planting had been done at the right time, that means indequacy of the rainfall may be due to poor planning based on peoples beliefs.

Rainfall analysis also help in transportation planning. The produce needs to be taken for marketing, therefore, an early estimate of the size of the yields would help to prepare storage, organize transport and arrangements of shifting the agriculture produce from one place to another. Poor transport system would result in wastage of agricultural produce. This type of planning can be done effectively if rainfall analysis is well established.

The other application of rainfall analysis is on water management. Long term average rainfall is very important for water conservation, although the variability of rainfall tend to be from year to year. Distribution of total rainfall in any given season is one of the most important aspects of rainfall analysis for agriculturalists and other water users.

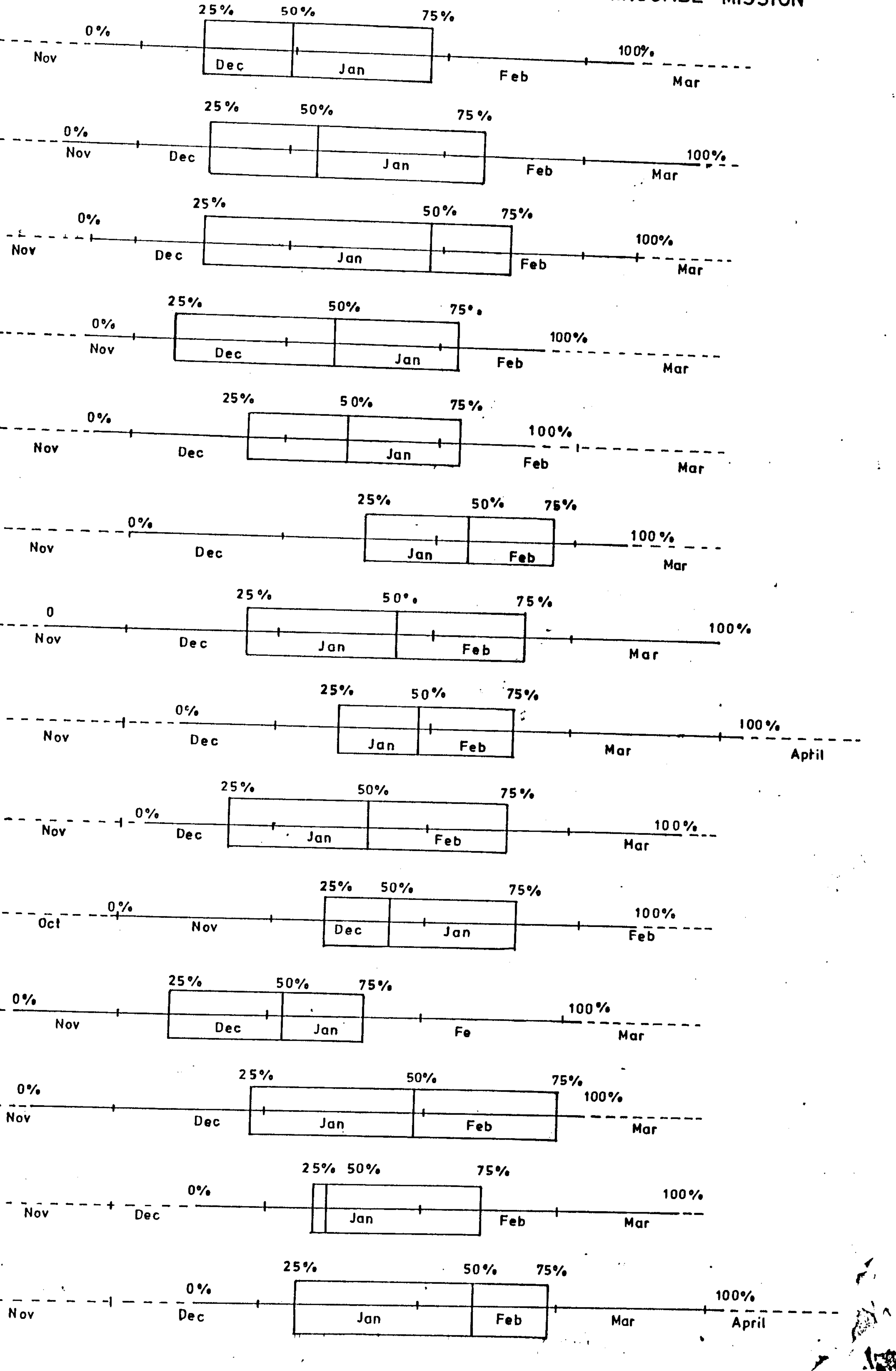
We are proposing some methods for the graphical illustration of annual rainfall patterns. These graphical methods are designed to give sufficient details of short term variabilities in heavier rainfall periods and to facilitate year comparisons of rainfall patterns.

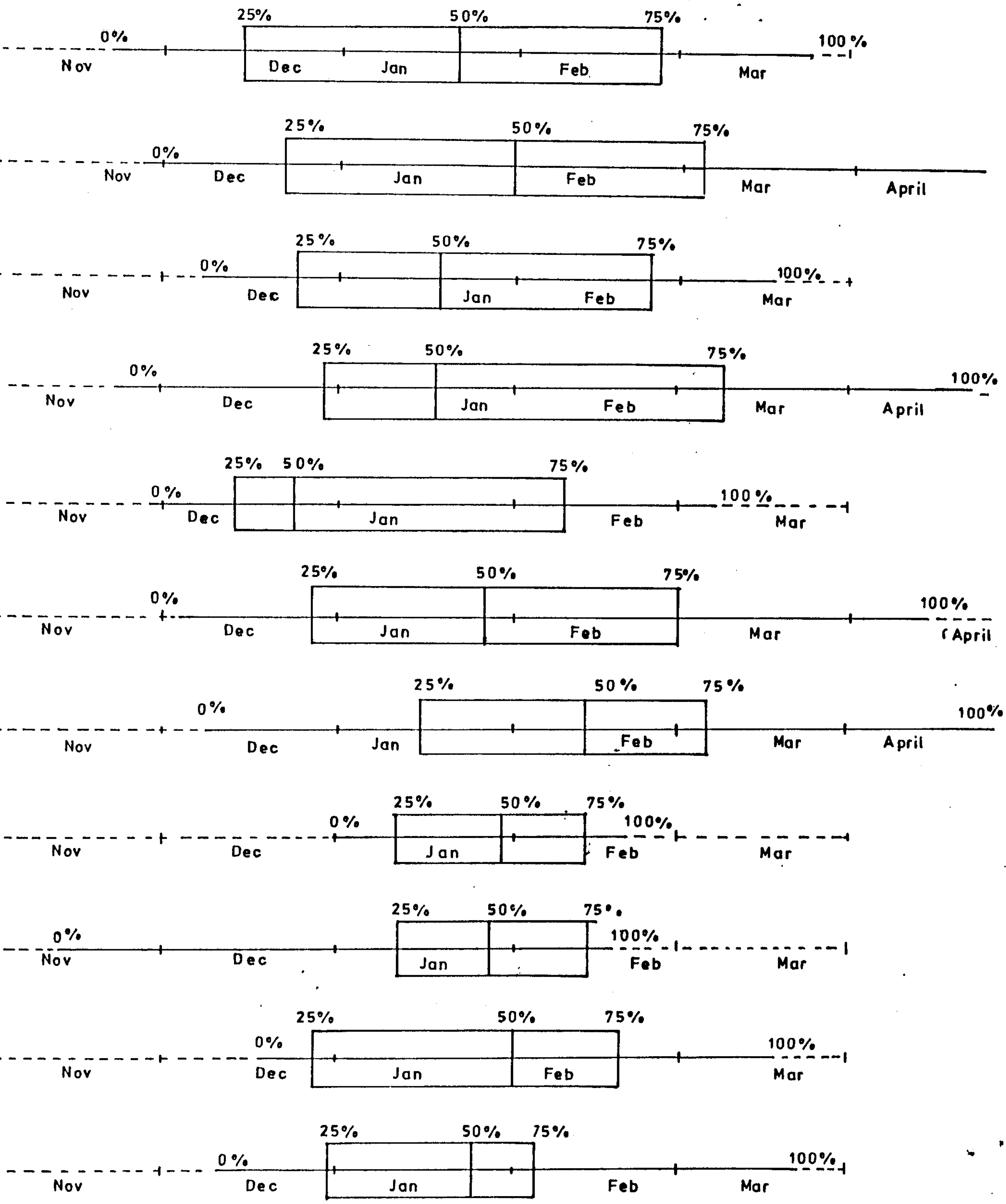
2.2. THE BOX PLOTS

Given the daily rainfall data at Chingombe weather station for the period 1960/1 to 1984/85, for each rainy season divide the total amount of rainfall into 25 per cent. This will give us four intervals with equal amount of rainfall in each interval. The length of each interval depends on the intensity of the total rainfall received in that period. If the length of the interval is short, that means 25% of the total rainfall occured within a short period, while if the length of the interval is long, that means the 25% of the total rainfall occured after a long time of rainfall. This simple form of showing the distribution of total rainfall allows us to make year to year comparisons. That is, we are able to see how much distributions of rainfall varies from year to year. The disadvantage of this form is that it doesn't show the actual variations on the amount of rainfall in each interval from year to year. Fig. 2.2. below illustrates the distributions of rainfall for each year using the box plots.

. 2·2

NFALL DISTRIBUTION FORM 1960/1-1984/5 SEASONS AT CHINGOMBE MISSION





The Rainy Season Period

Dry Season

2.3 DESIGN A HISTOGRAM

Using the same intervals established in section 2.2, distribution of total rainfall can be shown by constructing histograms in each interval by dividing each interval into seven equal parts (these equal parts are counted in days). We divide these intervals into seven equal parts since there is no single interval size which would work for the whole season due to variation of rainfall in any given season. Calculate the total amount of rainfall received in each of the seven equal parts of each interval, and then construct histograms in each interval by choosing column height $h_i Q_{jk}$ where $i = 1, \ldots\ldots, 7$, $j = 1, \ldots., 4$ and $k = 1, \ldots.., 25$ as total rainfall in each of the seven equal parts such that the total areas within each of the four intervals as indicated by the box plots (section 2.2) are equal. What we are trying to achieve by this construction is to find finer details of the intensity of the rainfall in each rainy season. Fig. 2.3 shows the constructed histograms showing the distribution of the total rainfall for the season.

The algorithm of constructing the histograms for the rainy season is as follows:

(a) Determine the beginning/end of the season using the definition given by the meteorological Department (countring days from 1st October, since there is no rainfall experienced in September for the period 1960/1 to 1984/5.

(b) Determine the days/dates when total rainfall equals to Q_{jk}, j = 1, ..., 4 k = 1, ..., 25

(c) Find the length of the Quartile

(Ljk, j = 1, ..., 4, k = 1, ..., 25)

(d) Divide each Quartile into seven equal parts:

i.e. $W_iQ_{jk} = L_{jk}$, i,...., 4 and j = 1, ..., 7, k 1, ..., 25

(e) Sum up the total rainfall in each of the seven equal parts (W_iQ_{jk});

i.e. h_iQ_{jk}

For each quartile construct a histogram with width = W_iQ_{jk} and column height = h_iQ_{jk}.

The algorithm of constructing the histograms for 1960/1 i.e. for k = 1 is as follows:-

The beginning of the season = 21st Nov. 1960 (52th day day from 1st October 1960).

The end of the season = 9th March 1961 (160th day from 1st Oct. 1960)

Q_1 occured on 13th December 1960 (74th day from 1st October, 1960)

Q_2 occured on 1st January 1961 (93rd day from 1st October 1960)

Q_3 occured on 28th January 1961 (120th day from 1st October 1960)

Q_4 = End of the season

L_{11} = 74-52+1 = 23 days; $W_iQ_{11} = {}^{23}/_{7} = 3\,{}^{3}/_{7}$ days, i = 1, ..., 7

Day	52	53	54	55	56	57	58	59	60	61	62	63	64	65	66	67	68	69	70	71	72	73	74
Rain fall	1.88	0.74	0.2	0	1.67	0	0	0	0.15	0	0	0	1.12	0	0	0	0.34	0	0	0	0.03	0.33	1.41

$^2/_7 \cdot {}^5/_7$ $\quad$ $^4/_7 \cdot {}^3/_7$ $\quad$ $^6/_7 \cdot {}^1/_7$ $\quad$ $^1/_7 \cdot {}^6/_7$ $\quad$ $^3/_7 \cdot {}^4/_7$ $\quad$ $^5/_7 \cdot {}^2/_7$

$h_1Q_{11} = 2.82$ $\quad$ $h_2Q_{11} = 1.67$ $\quad$ $h_3Q_{11} = 0.15$ $\quad$ $h_4Q_{11} = 0.12$ $\quad$ $h_5Q_{11} = 0.15$ $\quad$ $h_6Q_{11} = 0.19$ $\quad$ $h_7Q_{11} = 1.82$

For the first quartile, we construct a histogram with width 3 days and column heights as

$$h_1Q_{11} = 2.82 \text{ inches}$$
$$h_2Q_{11} = 1.67 \text{ inches}$$
$$h_3Q_{11} = 0.15 \text{ inches}$$
$$h_4Q_{11} = 0.12 \text{ inches}$$
$$h_5Q_{11} = 0.15 \text{ inches}$$
$$h_6Q_{11} = 0.19 \text{ inches}$$
$$h_7Q_{11} = 1.82 \text{ inches}$$

$L_{21} = 93 - 74 = 20$ days: $W_iQ_{21} = {}^{20}/_7 = 2\,{}^6/_7$ days; $i = 1, \ldots, 7$

Day	74	75	76	77	78	79	80	81	82	83	84	85	86	87	88	89	90	91	92	93
Rain fall	0.16	0	0	0	0.97	0	0	1.90	0	0.11	0	0	0	0	1.79	0.59	0.35	0.53	0.27	0.23

$^6/_7 \cdot {}^1/_7$ $\quad$ $^5/_7 \cdot {}^2/_7$ $\quad$ $^4/_7 \cdot {}^3/_7$ $\quad$ $^3/_7 \cdot {}^4/_7$ $\quad$ $^2/_7 \cdot {}^5/_7$ $\quad$ $^1/_7 \cdot {}^6/_7$

$h_1Q_{21} = 0.16$ $\quad$ $h_2Q_{21} = 0.97$ $\quad$ $h_3Q_{21} = 1.90$ $\quad$ $h_4Q_{21} = 0.11$ $\quad$ $h_5Q_{21} = 0.51$ $\quad$ $h_6Q_{21} = 2.30$ $\quad$ $h_7Q_{21} = 0.97$

For the second quartile, we construct histogram with width $2^6/_7$ days and column heights as:

$$h_1Q_{21} = 0.16 \text{ inches}$$
$$h_2Q_{21} = 0.97 \text{ inches}$$
$$h_3Q_{21} = 1.90 \text{ inches}$$
$$h_4Q_{21} = 0.11 \text{ inches}$$
$$h_5Q_{21} = 0.51 \text{ inches}$$
$$h_6Q_{21} = 2.3 \text{ inches}$$
$$h_7Q_{21} = 0.97 \text{ inches}$$

$l_{31} = 120 - 93 = 27$ days: $\quad W_iQ_{31} = {}^{27}/_7 = 3^6/_7$ days

Day	94	95	96	97	98	99	100	101	102	103	104	105	106	107	108	109
Rain fall	0.12	0	0.94	0	0.47	0.13	0	0	1.25	0.32	0	1.15	0	0.4	0.09	1.26

$^6/_7 \cdot {}^1/_7$ $\quad$ $^5/_7 \cdot {}^2/_7$ $\quad$ $^4/_7 \cdot {}^3/_7$ $\quad$ $^2/_7 \cdot {}^4/_7$

$h_1Q_{31} = 1.06$ $\quad$ $h_2Q_{31} = 0.6$ $\quad$ $h_3Q_{31} = 1.66$ $\quad$ $h_4Q_{31} = 1.09$

Day	110	111	112	113	114	115	116	117	118	119	120
Rain fall	0.03	0	0	0	0	0	0.25	0	1.37	0.05	0.09

$^2/_7 \cdot {}^5/_7$ $\quad$ $^1/_7 \cdot {}^6/_7$

$h_5Q_{31} = 0.75$ $\quad$ $h_6Q_{31} = 0.25$ $\quad$ $h_7Q_{31} = 1.51$

For the third quartile, we construct a histogram with width $3\,^6/_7$ days and column heights as

$$h_1Q_{31} = 1.06$$
$$h_2Q_{31} = 0.6$$
$$h_3Q_{31} = 1.66$$
$$h_4Q_{31} = 1.09$$
$$h_5Q_{31} = 0.75$$

$h_6 = 1.25$

$h_7 = 1.51$

$L_{41} = 160 - 120 + 1 = 41$ days; $W_i Q_{41} = {}^{41}/_{7} = 5\,{}^{6}/_{7}$ days

Day	120	121	122	123	124	125	126	127	128	129	130	131	132	133	134	135	136
Rain fall	0.15	0	0.15	0.04	0.07	0.14	0	0	0.3	0.06	0.46	0	0	0.12	0.44	0	0

${}^{6}/_{7} \cdot {}^{1}/_{7}$ ${}^{5}/_{7} \cdot {}^{2}/_{7}$

$h_1 Q_{41} = 0.53$ $h_2 Q_{41} = 0.81$ $h_3 Q_{41} = 0.67$

Day	137	138	139	140	141	142	143	144	145	146	147	148	149
Rain fall	0.19	0.63	0.31	0.13	0	0.27	0	0	0	0.57	0.53	0.42	0.04

${}^{4}/_{7} \cdot {}^{3}/_{7}$ ${}^{3}/_{7} \cdot {}^{4}/_{7}$ ${}^{2}/_{7} \cdot {}^{5}/_{7}$

$H_4 Q_{41} = 1.42$ $h_5 Q_{41} = 1.53$

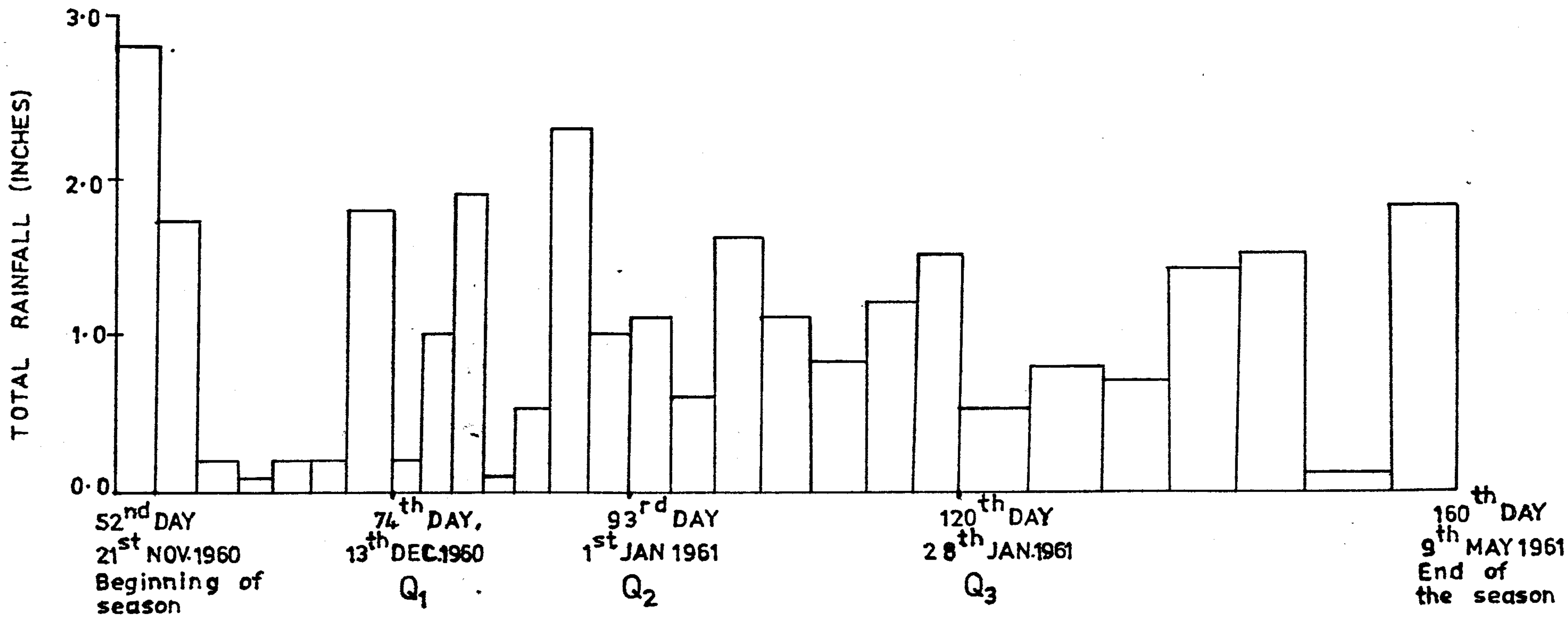
Day	150	151	152	153	154	155	156	157	158	159	160
Rain faill	0.06	0.03	0	0	0	0	0.2	0.86	0.04	0.39	0.34

$h_6 Q_{41} = 0.12$ ${}^{1}/_{7} \cdot {}^{6}/_{7}$ $h_7 Q_{41} = 1.83$

For the fourth quartile, we construct a histogram with width $5^6/_7$ days; and column heights as:

$$h_1Q_{41} = 0.53 \text{ inches}$$
$$h_2Q_{41} = 0.81 \text{ inches}$$
$$h_3Q_{41} = 0.67 \text{ inches}$$
$$h_4Q_{41} = 1.42 \text{ inches}$$
$$h_5Q_{41} = 1.53 \text{ inches}$$
$$h_6Q_{41} = 0.12 \text{ inches}$$
$$h_7Q_{41} = 1.83 \text{ inches}$$

RAINFALL DISTRIBUTION AT CHINGOMBE MISSION FOR 1960/1 RAINY SEASON



CHAPTER III

FUTURE SEASON PREDICTION

3.1 INTRODUCTION

Let X denote a variable associated with some aspects of a rainfall season. Since the daily rainfall data of Chingombe station for a period covering 25 seasons was available, there were 25 observed values of X. The objective is to predict the value of X for a subsequent season.

For the purpose of this project, it is necessary to limit the number of variables that may be used for the development of statistical procedures. But there are many variables which can be discussed with regards to the rainfall data. Due to vast variation in the daily rainfall both within and between seasons, it has to be acknowledged that the predictions may not always be particularly informative. This aspect will be discussed for each of the variables considered.

The approach will be as follows: Let $X_1, X_2, X_3, ---, X_{25}$ be the observed values of a variable X during 25 successive seasons. The first step is to check on whether the given data has a time trend. This was done using the run test for randomness and serial correlation methods.

3.2 THE RUN TEST FOR RANDOMNESS

Le $(X_1, X_2, X_3, \ldots\ldots., X_{25})$ be a random sample for some variable X. Given numbers X_1, X_2, X_3, ..., X_{25} in the sample space for X, consider the set P of all 25 - tuples containing exactly these 25 numbers. Since the sample is random, each 25 - Tuple in P has the same probability of being observed. Let M be the median of the numbers X_1, X_2, X_3, ...,X_{25}. Given a sample (X_1, X_2, X_3, ..., X_{25}) in P, we may replace X_i by A if $X_i > M$, replace X_i by B if $X_i < M$ and delete X_i if $X_i = M$

In the resulting sequence of A's and B's suppose there are a A's and a B's. These numbers a and b will be the same for each 25 - turple in p, since a is simply the number of the X_1, X_2, X_3, ..., X_{25} that exceed M, and similarly for b. Let R be the number of runs in the sequence of A's and B's obtained in this way, where a <u>run</u> is a sequence in which all terms are the same. How many different 25 - turples in p lead to the same arrangement of A's and B's ? Given the sequence of A's and B's, the probability distribution for R is given by WOLF F L (12) to be:-

$$p(R = 2k) = \frac{2\binom{a-1}{k-1}\binom{b-1}{k-1}}{\binom{a+b}{a}}$$

$$p(R = 2k + 1) = \frac{\binom{a-1}{k}\binom{b-1}{k-1} + \binom{b-1}{k}\binom{a-1}{k-1}}{\binom{a+b}{a}}$$

for k = 1,2, ..., min (a,b)

Using these results, it is possible to work out "critical values" for the number of runs. From table 1a on page 26

Given a and b, we find two numbers, say d and e, where d is the largest whole number such that $p(R \leq d) \leq 0.05$, and e is the smallest whole number such that $p(R \leq e) \geq 0.95$

Set up the hypothesis as follows:

H_o: The sample is random

H_o: The sample is non random

TABLE La: ONE SIDED 5 PER CENT CRITICAL VALUES FOR THE NUMBER OF RUNS

b \ a	5	6	7	8	9	10	15	20
5	3,8	3,9	3,9	3,10	4,10	4,10	5,11	5,11
6		3,10	4,10	4,11	4,11	5,11	6,13	6,13
7			4,11	4,12	5,12	5,11	6,14	7,14
8				5,12	5,13	6,13	7.15	8.16
9					6.13	6,14	8,16	9,17
10						6,15	8,17	9,18
15							11,20	12,23
20								15,26

TABLE 2(a) THE RUN TEST FOR RANDOMNESS

VARIABLE	MEDIAN	NUMBER OF A's	NUMBER OF B's	RUNS	SIGNIFICANCE
X_1	54	12	10	10	non-significance
X_2	41	12	12	17	non-significance
X_3	82 mm (3.23")	12	12	12	non-significance
X_4	341 mm (13.43")	12	12	6	non-significance
Y_1	184	12	11	12	non-significance
Y_2	882 mm (34.72")	12	12	12	non-significance

3.3 SERIAL CORRELATION

In addition to the run test for randomness applied to each variable, a test of randomness of a variable X can also be established by serial correlation. A serial correlation of about 0.4 or more is necessary for the sample to be regarded as significantly different from zero.

Table 2(b) below gives summary of the serial correlation of the observed values of the the variables considered.

TABLE 2(b) SERIAL CORRLATION

VARIABLE	LAG 1	LAG 2	LAG 3	LAG 4	LAG 5	SIGNIFICANCE
X_1	0.160639	-0.097954	-0.282779	0.049475	0.284050	non-significance
X_2	0.356002	-0.371950	0.300705	0.117869	-0.254400	non-significance
X_3	0.030567	0.046597	0.102282	-0.32999	-0.266928	non-significance
X_4	0.038702	0.059607	-0.339631	-0.083138	-0.070613	non-significance
Y_1	-0.025208	0.093876	-0.032756	0.168323	-0.153630	non-significance
Y_2	-0.155205	0.053920	0.146274	-0.130578	-0.102391	non-significance

From the results of the RUN TEST FOR RANDOMNESS AND SERIAL CORRELATION there were no grounds for rejecting the assumption that the observed X values were random observations from some probability distribution. Having established the randomness of the X values, it is necessary to consider ways of predicting X_{n+1}, that is, the value of X for the next season.

In particular $p_c = p(X_n+1 \leq C)$ will be estimated for some specified values of C. In this case C is the number of days from the origin, the origin used here being 23rd September or the rainfall measured in millimetre. Distribution - free estimation is straight forward. Suppose that the choice of C is in the interval $[X_r, X_{r+1}]$, that is, $[X_r \leq C \leq X_{r+1}]$, where X_r denotes the r^{th} ordered observation. Then, clearly an unbiased estimate of $\hat{p}_c$ is $p_c = \frac{r}{n}$ for any C in that interval. Furthermore, using cumulative binomial probabilities, it is possible to determine iteratively confidence limits for p_c. In this case n=25, the standard error of $\hat{p}_c$ will be about $\sqrt{\frac{1}{4n}} \simeq 0.1$ so that for example, a 95 per cent confidence interval for $\hat{p}_c$ will be of width which is approximately $2 \times 2 \times 0.1 = 0.4$. This width is clearly too much for the estimate of p_c to be informative. In view of this and the remarks made above that p_c will be the same for a whole range of values of C, the results of this method have not *been reproduced here.*

3.4 TRUNCATED EXPONENTIAL DISTRIBUTION

Now consider setting up a distributional model for X. The

Nature of the type of X considered here does not suggest any obvious choice as far as is known, since no extensive study has been made of such a random variable. Since our sample size of 25 is rather small, any goodness of fit test applied to the data is most unlikely to reject whatever model is postulated. *To provide some flexibility, a distribution* model involving at least two unknown parameters is preferable. Two - parameter distributions that might be worth trying include, the Lognormal, the Truncated exponential and the Weibull. A much large sample than what is available in this project would be necessary to determine which of these is the most appropriate. For analytic simplicity, suppose that X has the Truncated exponential distribution of density:-

$$f(x) = \frac{1}{\alpha}e^{-(x-\beta)/\alpha} \quad , \ x>\beta$$

where α and β are unknown parameters. Using results derived by Lawless J.F. [9] it may be shown that given the sample values $X_1, X_2, \ldots, X_n$ the predictive distribution of X is such that:

$$p(X \leq C) = 1 - \frac{n}{n+1}\left(1 + \frac{C - x_1'}{U}\right) - (n-1) \quad , \text{ for } C \leq X$$

where x_1' is the smallest sample value and

$$U = \sum_{i=1}^{n} x_i - nx_1'$$

Thus, for any given p, $p(X \leq C) \geq \frac{1}{n+1}$, we find that

$$C = x_1' + U\left\{\left[\frac{n}{(n\ 1)(1-p)}\right]^{\frac{1}{n+1}} - 1\right\}$$

which is an upper 100p percent bound for the predicted value of X. Applying these results to the various X's and Y's under consideration, table 3 below gives the 100p percent upper bounds for the predicted X and Y for selected values of p. When appropriate, the value of C (in days) is converted into a date.

TABLE 3: 100p% Upper Bounds for the predicted X for selected Values of P for 1985/6 session

Variable	P = P(X C)	0.5	0.6	0.7	0.8	0.9	0.95
X_1	C	47	56	67	84	133	143
	DATE	9/11/85	18/11/85	29/11/85	16/12/85	3/2/86	13/2/86
X_2	C	35	39	46	55	70	122
	DATE	28/10/85	1/11/85	8/11/85	17/11/85	2/12/85	24/1/86
X_3	C	28mm	101mm	126mm	161mm	224mm	288mm
X_4	C	154mm	316mm	372mm	452mm	592mm	735mm
Y_1	C	173	185	198	219	254	309
	DATE	14/3/86	26/3/86	8/4/86	29/4/86	3/6/86	28/7/86
Y_2	C	537mm	885mm	1006mm	1178mm	1479mm	1789mm

In the above table, the C is an upper 100P% bound for the value of X in the next season given by the Truncated exponential distribution. Thus for example, there is a 50% chance that the next rainy season (X_1) will start within 47 days from 23rd September 1985 (that is, before 9th November 1985) and 90% chance that it will start within 133 days (that is before 3rd February 1986).

Similarly, there is a 50% chance that the first day of at least two consecutive days with rainfall and the accumulated rainfall is 10mm or more. (X_2) occurs within 35 days (that is by 28th October, 1985) and a 90% chance that it occurs within 70 days (that is, by 2nd December, 1985). Further, there is a 50% chance that (i) the total rainfall up to 30th November 1985 (X_3) will be less than 28mm, (ii) the total rainfall up to 31 November 1985 (X_4) will be less than 154mm (iii) the total rainfall (Y_2) will be less than 537mm.

The results displayed in Table 3 are not particularly informative for agriculturalists. It is possible that an alternative distribution model, such as the Lognormal or the Weibull, would have been a better choice. Unfortunately, time has not permitted such possibilities to be investigated. As mentioned earlier, the advantage of the Truncated exponential over the other possible distribution was that an explicit expression can be obtained for the prediction bounds.

CHAPTER IV

WITHIN SEASON PREDICTIONS

4.1 INTRODUCTION

Given data on the rainfall in the early part of a season, we now consider the problem of making predictions about certain aspects of the rainfall that will occur in the remaining part of the season. In particular, we shall assume that the values of some or all of X_1, X_2, X_3 and X_4 are available in a certain season and we shall predict the values of one or more of Y_1 and Y_2 (Y's as defined in section (1.1). An abvious method for achieving this is by means of linear regression analysis.

Regression has many uses. The objectives may be to determine if there is a relationship between Y and X, to study the shape of the relationship, and to think about the reasons for the relationship. Prediction of Y from X may be the goal. Often the purpose is to estimate the change in Y to be expected from a given increase in X in cases where we are convinced that changes in Y are mainly caused by changes in X. If Y represents the dependent variable and X the independent variable, then the equation of a straight line relating to these two variables called a model is:

$$Y = \beta_o + \beta_1 X + \varepsilon$$

Where β_o is the intercept, β_1 is the slope and ε is the error. Let the difference between the observed value of

Y and the straight line ($\beta_o + \beta_1 X$) be an error ε. It is convenient to think of ε as a statistical Error; that is, a device that accounts for the failure of the model to fit the data exactly. The error may be made up of the effects of the variations of variables, measurement errors or change in weather due to the movement of the earth.

The two basic things in this analysis which are important are:

(a) Estimating Y corresponding to the regressor X values

(b) The adequacy of the model

This out come of adequacy checking may indicate either that the model is reasonable or that the original fit must be modified. This is an iterative procedure in which data lead to a model and a fit of the model if the data is produced. A regression model does not imply a cause and effect relationship between variables. Even though a strong empirical relationship may exist between two or more variables, this can not be considered as evidence that the regressor variables and the response are related in a cause-effect manner. To establish

---

P.S.: *The contents of this chapter have been presented in Cosmex-89: An International Conference on Stochastic Methods in Experimental Sciences (7-14, september 1989) Wroclaw, Poland in a joint paper with Dr. J. Prakash entitled "Analysis of Daily Rainfall Data in Zambia" which has been accepted for publication in the Conference Proceedings.*

the relationship between the regressor and the resposes must have a basis outside the sample data, say, a relationship suggested by theoretical considerations. In other words, regression analysis can aid in confirming a cause effect relationship, but it can not be the sole basis of such a claim. One point worth noting is that it is usually more important to gain insight and understanding concerning the system generating the data.

Set up = number of parameters and n = number of observations. Then MS_E and t have (n-p) degrees of freedom.

Consider the linear regression model

$$Y = \beta_o + \beta_1 X_1 + \beta_2 X_2 + \cdots\cdots + \beta_n X_n + \varepsilon_i$$

Where $X_1, X_2, \ldots, X$ are the regressor (or Independent) variables and ε is an error term. Consider fitting this model to observed values $(Y_i : X_{1i}, X_{2i}, \ldots, X_{\ i})$, i = 1, 2, ..., k then

$$Y_i = \beta_o + \beta_1 X_{1i} + \cdots + \beta_n X_{\ i} + \varepsilon_i, \quad i = 1, 2 \ldots, k$$

We shall assume that the ε_i are Independent and normally distrubuted with mean zero and constant variance σ^2. In matrix notation, the above k equations, become

$$Y = XB + E$$

Where $Y = (Y_1, Y_2 \ldots, Y_k)$, $B = [\beta_o, \beta_1, \ldots, \beta_n]$ and $E = (\varepsilon_1, \varepsilon_2, \ldots \varepsilon_R)$ are column vectors and

$$X = \begin{bmatrix} 1 & X_{11} & \cdots & X_{n1} \\ \cdot & \cdot & & \cdot \\ \cdot & \cdot & & \cdot \\ \cdot & \cdot & & \cdot \\ 1 & X_{1k} & \cdots & X_{nk} \end{bmatrix}$$

Then, the least square estimate of B is given by:

$$\hat{\beta} = (X'X)^{-1}XY$$

and for any specified $X_o = [1, X_1, \ldots, X_n]$, the least squares estimate of the observed values Y_o of Y is $Y_o = X_o'\hat{\beta}$ where sampling variance is given by:

$$VAR(\hat{Y}_o) = [1 + X_o'(X'X)^{-1}X_o] \; \sigma 2$$

Furthermore, the 100 $(1 - \alpha)$% predicted interval for Y_o has the limits

$$Y_o \pm t\alpha/_2, \; k\text{-}n \sqrt{MS_E \quad 1 = X_o'(X'X)^{-1}X_o}$$

Where $$MS_E = \frac{(Y - X'\hat{\beta})\;(Y - X'\;\hat{\beta})}{k\text{-}n\text{-}1}$$

4.2 Hypothesis testing on the slope

We are often interested in testing hypothesis and constructing prediction intervals about the model parameter. These procedures require that we make the additional assumption that the model errors are normally distributed. Thus the complete assumptions are that the errors are normally distributed and independantly distributed with mean 0 and varience σ^2, abbreviated $(0,\sigma^2)$.

A very important special case is:

$$H_o: \quad \beta = 0$$

$$H_1: \quad \beta \neq 0$$

This hypothesis relate to the significance of regression. Failing to reject H_o: $\beta = 0$ implies that there is no linear relationship between X and Y. This situation is illustrated in Fig 4a. Note that this may imply either that X is of little value in explaining the variation in Y and that the best estimator of Y for any X is $Y = \bar{Y}$ (Fig. 4a (i) or that the true relationship between X and Y is not linear (Fig 4a(ii). Alternatively, if H_o : B = 0 is rejected, this implies that X is of little value in explaining the variability in Y. This is illustrated in Fig 4a. However, rejecting H_o: B = 0 could mean either that the straight line mode is adequate (Fig 4b(i) or that even though there is a linear effect of X, better results could be obtained with the addition of higher order polynomial terms in X (Fig 4b(ii).

Fig 4a

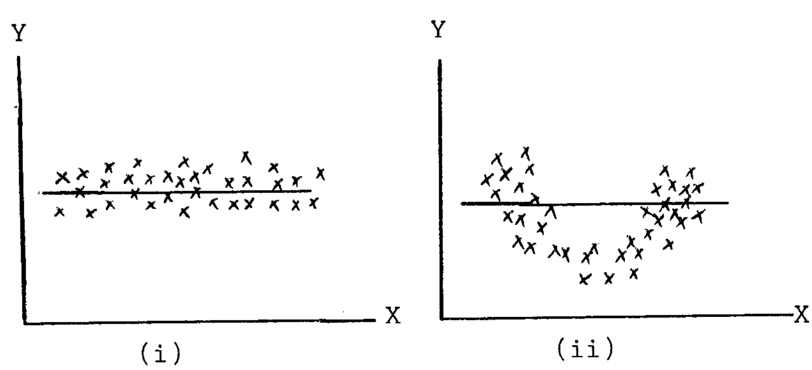
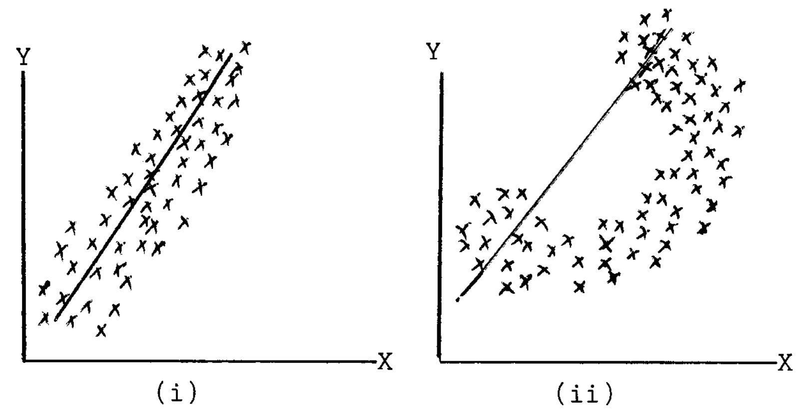


Fig 4b



To test the hypothesis H_o : B = 0, an analysis of variance procedure is used.

The test statistics is:

$$F_o = \frac{SS_{R/1}}{SS_{E/(n-2)}} = \frac{MS_R}{MS_E}, \text{ where}$$

MS_R is the regression mean square and MS_E is the residual mean square. The test for significance of regression may also be performed using the t - test equation with $\hat{\beta}_1$ 0, say

$$t_o = \frac{\hat{\beta}_1}{MS_{E/S_{xx}}}$$

$$t_o^2 = \frac{\hat{\beta}_1^2 S_{xx}}{MS_E} = \frac{\beta_1 S_{xx}}{MS_E} = \frac{MS_R}{MS_E}$$

Although the t-test for H_o: β_1 = 0 is equivalent to the F-tests, the t-test is somewhat more adaptable, as it could be used for one-sided alternative hypothesis (either H_i: β_1<0∞ H_1: β_1>0), while the F-test considers only the two-sided alternative. The inability to show that the slope is not statistically different from zero may not necessarily mean that Y and X are unrelated. It may mean that our ability to detect this relationship has been obscured by the variance of the measurement process or that the range of values of X is inappropriate. A great deal of non statistical evidence and knowledge of the subject matter in the field is required to conclude that $\beta_1 = 0$.

4.3 ADEQUACY OF MODELS

It is important to have some checks that the assumptions made in a regression analysis are valid for the problem under consideration. Consider the residuals (errors)

$$\varepsilon_i = Y_i - \hat{Y}_i, \qquad i = 1, 2, \ldots\ldots, n$$

where $Y_i = X_i'\hat{\beta}$, with $X_i' = [1, X_{ii}, \ldots, X_{iii}]$

It is necessary to consider the validity of the assumptions used on the linear model before it is used. Gross violations of the assumptions may yield an unstable model, in the sense that a different sample could lead to a total different model with different conclusions. The t or F statistics or R^2 cannot detect departures from the underlying assumptions. We shall assume that given $X_i = (X_{1i}, X_{2i} \ldots\ldots X_{ni})$, the observed values of Y_i are such that $Y_i = \beta_. + \beta_1 X_i + \varepsilon_i$, $i = 1, 2, \ldots., n$. Where the ε_1 are independent and normally distributed with variance σ^2. For any X_i, the value of Y on the least square line is: $\hat{Y}_i = \hat{\beta}_o + \hat{\beta}_i X_i$ and the residuals are: $\varepsilon_i = Y_i - \hat{Y}_i$, $i = 2, \ldots\ldots, n$

which is the deviations between the data and the fitted values. There are many wasy of assessing the adequacy of the model. The usual checking methods are as follows:

(a) NORMAL PROBABILITY PLOT:

This is designed so that the cumulative normal distributions will plot as a straight line, when the residuals are ranked in increasing order. Substantial departures from a straight line indicate that the distribution is not normal.

(b) THE PLOT OF THE STANDARDIZED RESIDUALS

The Standardized residuals are useful for detecting departures from normality if the errors are normally distributed. Approximately 95 per cent of these standardized residuals should fall between -2 and +2. Substantial deviations from these limits indicates potential violation of the normality assumptions.

(c) PLOT OF STANDARDIZED RESIDUALS AGAIST STANDARDIZED PREDICTED Y_i:

If all the standardized observations are contained between -2 and +2, then there are no obvious model defects. A ventical band of residuals indicate that a long-term time effect is not influencing the data (or if it is, has somehow been accounted for by an X-variable also subject to a time effect). This vertical band would mean that the residuals have a constant variance; otherwise a weighted least squares analysis should be used, or quadratic terms in time should be included in the model.

(d) COEFFICIENT OF DETERMINATION

The quantity $R^2 = \frac{SYY - SSE}{SYY}$, where $SYY = \Sigma(Y_i - \bar{Y})^2$ and $SSE = \Sigma_{\delta i}{}^2$ is called the coefficient of determination. SYY is a measure of the variability in Y without considering the effect of the regressor variable X, and SSE is a measure of the variability in Y remaining after X has been considered. R^2 is called the proportions of variations explained by the regressor. Values of R^2 that are close to 1 imply that most of the variability in Y is explained by the regressor model. This statistics R^2 should be used with caution since it is always possible to make R^2 large by adding enough terms to the model.

4.4 PREDICTING Y_1 (END OF RAINY SEASON)

In this section we shall consider the prediction of the end of the rainy season having observed the values of X_1, X_2, X_3 and X_4 for the season 1985/86. For the 1985/86 season the actual observed values were:

$X_1 = 41$mm, $X_2 = 37$mm, $X_3 = 2.9$ inches $X_4 = 7.68$ inches

and we shall use these values to demonstrate the theoretical results.

4.4.1 Simple Linear Regression

An important application of the regression model is the prediction of a new observation of Y corresponding to a specified level of the regressor variable X. If X_o is the value of the regressor variable of interest then:

$$\hat{Y}_1 = \hat{\beta}_o + \hat{\beta}_1 X_o$$

where X_o is one of X_1, X_2, X_3, X_4 (observed values for the first part of the 1985/86 rainy season). Y_1 is the point estimate of the new value of the response Y_1.

The 100(1-α) per cent prediction interval on a future observation at X_1 is:

$$Y_1 \pm t\alpha/_2, \; k-n \sqrt{MS_E\left[1 + \frac{1}{n} + \frac{X_o - \bar{X})^2}{S_{xx}}\right]}$$

This prediction interval is of minimum width at $X_o = \bar{X}$ and wides as $|X_. - \bar{X}|$ increase. We observe that the prediction interval at X_o is always wider than the confidence interval at $\bar{X}$, because the predicted interval depends on both the error from the fitted model, and the errors associated with future observations. Here it is important to look at the relationship between Y_1 = The end of the rainy season and X_1, X_2, X_3 and X_4 as observed values for the 1985/86 rainy season. Anova. table continued.

TABLE 4.4b

Dependant Variable Y_1					
X_1	DF	SS	MS	F	Signif. F
Regression	1	1322.89610	1322.89610	3.36694	0.0795
Residual	23	9036.86390	392.90713		
Non Significant					
X_2					
Regression	1	179.18544	179.18514	0.40482	0.5309
Residual	23	10180.57486	442.63369		
Non Significant					
X_3					
Regression	1	142.65869	142.65869	0.32114	0.5764
Residual	23	10217.10131	444.22180		
Non Significant					
X_4					
Regression	1	170.26430	170.26430	0.38433	0.5414
Residual	23	10189.49570	443.02155		
Non Significant					

TABLE 4.4b

ANOVA

Dependant variable Y_1					
X_1 X_2	DF	SS	MS	F	Signif F
Regression Residual	2 22	1813.09750 8546.66250	906.54875 388.48466	2.33355	0.1205
Non Significant					
X_1, X_3					
Regression Residual	2 22	2379.63443 7980.12557	1189.81722 362.73298	3.28015	0.0567
Non Significant X_1, X_4					
Regression Residual	2 22	2119.09196 8240.66804	1059.54598 374.57582	2.82866	0.0807
Non Significant X_2, X_3					
Regression Residual	2 22	296.95708 10062.80292	148.47854 457.40013	0.32461	0.7262
Non Significant X_2, X_4					
Regression Residual	2 22	334.88553 10024.87447	167.44277 455.67611	0.3646	0.6967
Non Significant X_3, X_4					

TABLE 4.4b CONTINUED

ANOVA

X_1, X_2	DF	SS	MS	F	Signif F
Regression	*2*	*190.59092*	*95.29546*	*0.20616*	*0.8153*
Residual	22	462.23496			
Non Significant					
X_1, X_2, X_3					
Regression	3	2903.82379	967.94126	2.72625	0.0699
Residual	21	7455.93621	355.04458		
Non Significant					
X_1, X_2, X_3					
Regression	3	2506.28397	835.42799	2.23391	0.1141
Residual	21	7853.47603	373.97505		
Non Significant					
X_2, X_3, X_4					
Regression	3	347.61364	115.87121	0.24303	0.8653
Residual	21	10012.14636	476.7887		
Non Significant					
X_1, X_2, X_3, X_4					
Regression	4	3049.64941	762.41235	2.08591	0.1207
Residual	20	7310.11059	365.50553		

Non Significant

4.4.2 MULTIPLE LINEAR REGRESSION

The results of the multiple linear regression analysis are the extention of those discussed in section 3.3.1 on simple regression. In general the response Y may be related to k regressor variables.

Consider the model

$$Y = \beta_o + \beta_1 X_1 + \dots . + \beta_k X_k +$$

With error terms uncorrelated, normally distributed with mean zero and constant variance σ^2. Multiple linear Regression models are often used as approximating functions. That is, the true relationship between E [Y] and X_1, X_2, ..., X_k is unknown, but over certain ranges of the regressor variables the model is an adequage approximations. In general, any regression model that is linear in the parameters (the $\beta_s^{'}$) is linear Regression model, regardless of the shape of the surface that it generates.

For multiple regression analysis, although the response variable seem to be significantly related to the regressor variables from the analysis of varieance tables, it is possible that only one regressor variable contributes significantly to the relationship of Y and X. To establish whether the regressor variable contribute significantly to the relation of Y and X, we simply test H_o: $\beta = 0$ for each regressor coefficient. If $\beta \neq 0$, then the multiple regression analysis using this variable would not give a reliable result for predication purposes. It will be evident from the tests outlined below which regressor variables are significantly related to Y for multiple regression analysis purposes.

Multiple regression model can also be used to predict future observations on Y corresponding to particular values of the regressor variable, for example, X_1, X_2, X_3,X_4. If $X_6' = 1$, X_1, X_2, X_3,X_4, then a point estimate of the future observation Y_o at the point X_o,

$$\hat{Y} = X_o'\hat{\beta} \qquad \text{(A)}$$

$$Y = \hat{\beta}_o + \sum_{i=1}^{4} + \hat{\beta}_i X_i$$

The vector of fitted values Y corresponding to the observed values Y is:

$$\hat{Y} = X\hat{\beta}$$

$$\hat{Y} = X(X'X)^{-1}X'Y$$

$$\hat{Y} = HY$$

The n x matrix $H = X(X'X)^{-1}X'$ is usually called the "hat" matrix, because it maps are the vector of observed values into a vector of fitted values. The hat matrix and its properties play a central role in regression analysis.

To construct a predicted or predicted interval on the response at a particular point, such as X_1, X_2, X_3, X_4.

Define the vector X_o as

$$X_o = \begin{bmatrix} 1 \\ X_1 \\ X_2 \\ X_3 \\ X_4 \end{bmatrix}$$

(A) is the fitted value at this point, which is an unbiased estimator of Y since $E[Y] \equiv \hat{Y} = X_o' \hat{\beta}$ and the variance of Y is

$$V(\hat{Y}) = MS_E X_o'(X'X)^{-1}X_o$$

Therefore, a 100 (1 $-\alpha$) percent predicted interval for the future observation is:

$$\hat{Y} - t_{\alpha/2,\ n-p} \sqrt{MS_E \quad (1+X_o'(X'X)^{-1}X_o)} \leq Y \leq \hat{Y} + t_{\alpha/2,\ n-p} \sqrt{MS_E(1+X_o'(X'X)^{-1}X_o)}$$

4.5. PREDICTING Y_2 (TOTAL RAINFALL FOR THE SEASON)

In this section we consider the prediction of the total rainfall for 1985/86 rainy season having observed the values X_1, X_2, X_3 and X_4 for that season. The actual observed values for 1985/86 rainy season as derived from Appendix D are: X_1 = 41, X_2 = 37, X_3 = 2.9" and X_4 = 7.7" (" = inches) and the theoretical results are:

TABLE 4.5a

LEAST SQUARES ESTIMATES FOR ALL POSSIBLE REGRESSIONS (RAINFALL DATA)

DEPENDENT VARIABLE Y_2						
VARIABLE	β_o	β_1	β_2	β_3	β_4	F
X_1	38.11537	-.010839				1.36308
X_2	19.48357		0.29517			4.85680*
X_3	25.28937			1.63845		6.00553*
X_4	22.22603				0.73620	5.15837*
X_1 X_2	24.63462	-0.06881	0.27128			2.67742
X_1 X_3	26.40394	-0.01495		1.58400		2.88684
X_1 X_4	25.55278	-0.04267			0.67792	2.59226
X_2 X_3	14.70821		0.26993	1.52233		5.97045*
X_2 X_4	11.06527		0.28254		0.70629	5.73357 **
X_3 X_4	22.37749			1.10177	0.39202	3.41257
X_1 X_2 X_3	12.32716	0.02772	0.27794	1.61985		3.84532*
X_1 X_3 X_4	22.91445	-6.92579 E-03		1.08032	0.38926	2.17377
X_1 X_3 X_4	11.61853		0.27190	0.96626	0.40557	4.35676*
X_2 X_2 X_3 X_4	8.31311	0.03712	0.28270	1.07586	0.42089	3.17669*

* = Significant at 95%

** = Significant at 99%

4.5.1 SIMPLE LINEAR REGRESSION

Consider the model

$$\hat{Y}_2 = \hat{\beta}_o + \hat{\beta}_1 X_o$$

where X_o is one of X_1, X_2, X_3 or X_4 (observed values for 1985/86 season).

4.5.1.1 Predicting Y_2 from X_2

The scater plot on Fig 4.5 shows that the total rainfall of the rainy season is significantly related to the first day of at least two consecutive days with rainfall and the accumulated rainfall is 10 mm or more moresured from 1st October. Table 4.5(a) shows the analysis of variance for Y_2 against X_2. From this ANOVA table significance F = 0.0378 0.05, which means we reject H_o at 95% level of confidence. Table 4.5(b) shows the t - values of each variable. Fig 4.5(b) and 4.5(c) shows the normal probability ($\hat{p}$-p) plot and standardized scatter plot respectively. From these plots it is evident that our model is adequate.

TABLE 4.5(a)

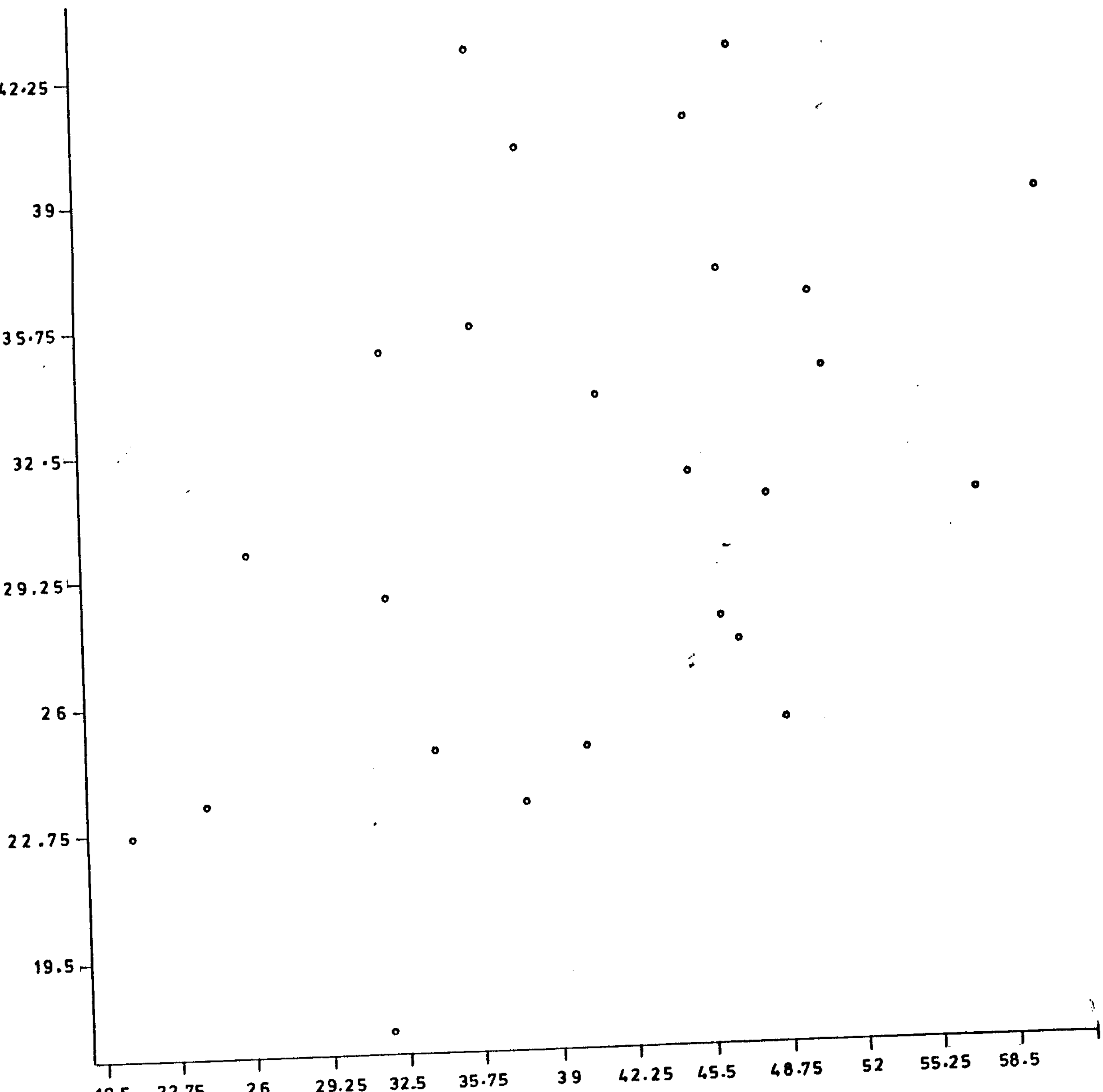
ANALYSIS OF VARIANCE TABLE

	DF	SUM OF SQUARES	MEAN SQUARE
Regression	1	203.82029	203.82029
Residual	23	965.21805	41.96600
F = 4.85680		Significance F = 0.0378	

TABLE 4.5(b)

Variable	β_o	S.E.	95% Confidence Interval of	T-value	Sig. t
X_2	0.29517	0.13394	0.01810 - 0.57224	2.204	0.0378*

Fig 4.5(b) PLOT OF Y2 WITH X2



25 Cases plotted

Fig 4.5 (b)

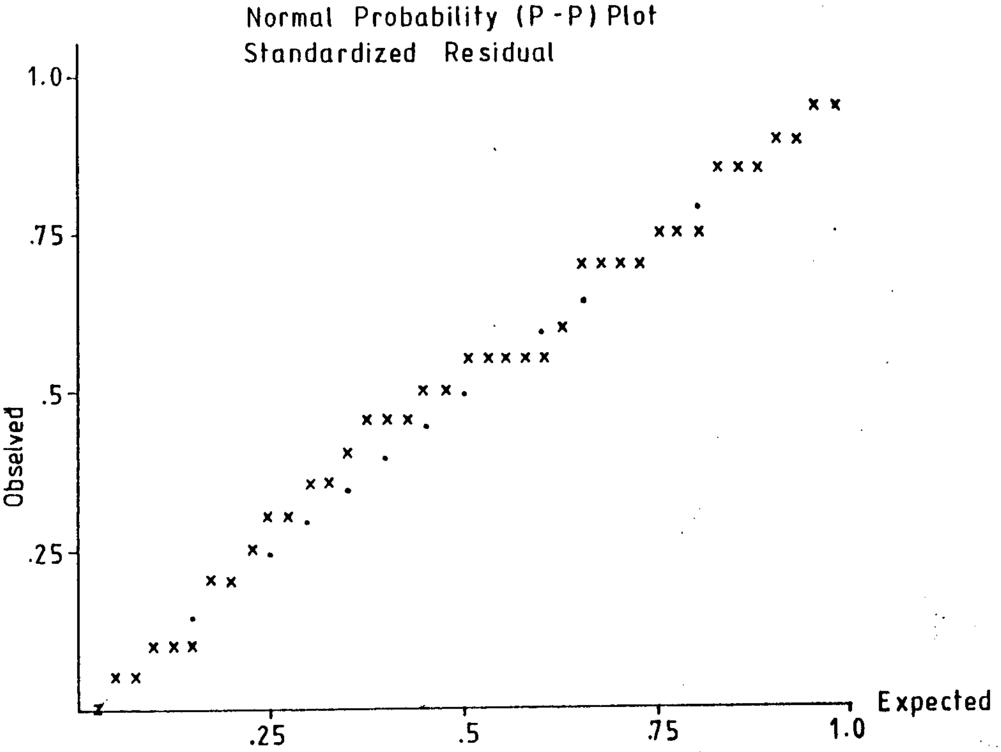
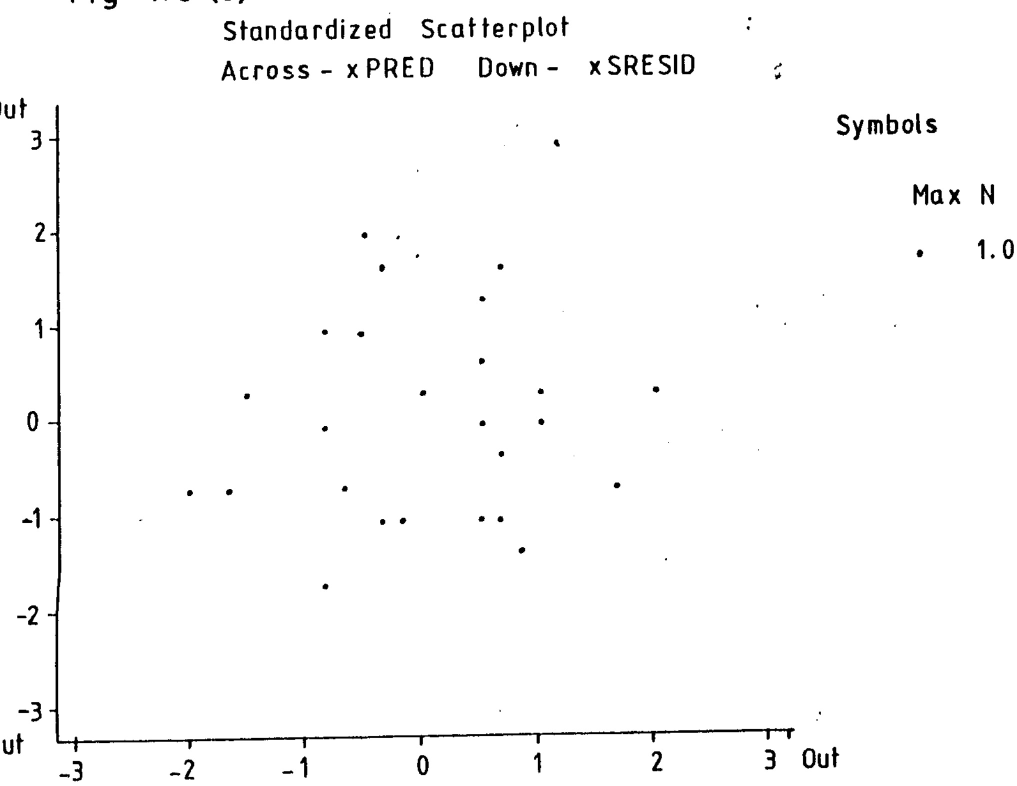


Fig 4.5 (c)



The Equation is:

$$\hat{Y}_2 = 19.48357 + 0.29517X_2$$

$$\hat{Y}_2 = 19.48357 + 0.29517 \times 37$$

$$\hat{Y}_2 = 30.4 \text{ inches } (772.28\text{mm})$$

That means, given that the first day when accumulation rainfall is greater than or equal to 10mm (X_2) then the total rainfall for the rainy season would be 30.4 inches (772.28mm).

A 100(1 $-\alpha$) percent prediction interval on future observation at X_2 is:

$$\hat{Y}_2 \pm t_{\alpha/2,\ k-n} \sqrt{MS_E \left[1 + \frac{1}{n} + \frac{X_o - \bar{X})^2}{S_{xx}} \right]}$$

$$\text{Where } S_{xx} = \sum_{i=1}^{4} x_i^2 - \frac{(\sum_{i=1}^{n} X_i)^2}{n} = \sum_{i=1}^{n} (x_i - \bar{X})^2$$

$$\hat{Y}_2 \pm t \quad 0.025,23 \sqrt{41.966 \left[1 + \frac{1}{25} + \frac{(37 - 40.84)^2}{2339} \right]}$$

$30.4 \pm 2.069 \times 6.6$

$\underline{30.4 \pm 13.7}$

That means, we are 95% confident that the total rainfall for 1985/6 rainy season will be between 16.7 inches (424.18mm) and 44.1 inches (1120mm)

4.5.1.2 Predicting Y_2 from X_2

The scatter plot on Fig 4.5d shows that the total rainfall of the rainy season is significantly related to the total rainfall up to 30^{th} November. Table 4.2a shows the ANOVA for Y_2 against X_3. From the ANOVA table significance F 0.0223. Figs 4.5d, 4.5e shows the normal probability plot and standardized scatter plot respectively.

TABLE 4.5c

ANALYSIS OF VARIANCE

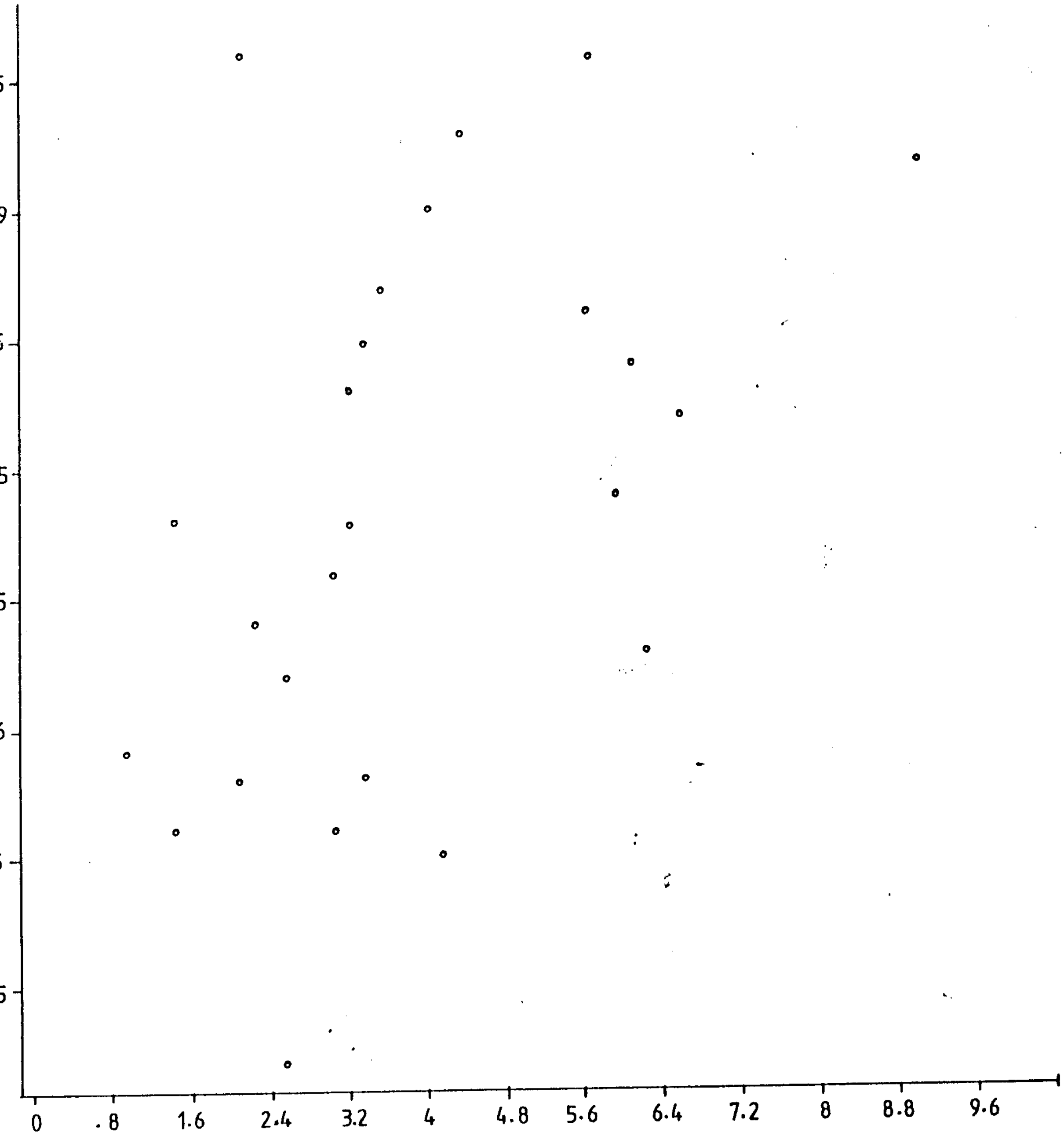
	DF	SUM OF SQUARES	MEAN SQUARE
Regression	1	242.04668	242.04668
Residual	23	926.99166	40.30399
F = 6.00553		Significance F = 0.0223*	

TABLE 4.5d

Variable	β_o	S.E. β	95% confidence interval of β	t-value	Sig t
X_3	1.63845	0.66858	0.25537 - 3.02152	2.451	0.0223*

Fig 4.5d

PLOT OF Y2 WITH X3



25 Cases plotted

Fig. 4.5d

Normal Probability (P-P) Plot Standardized Residual

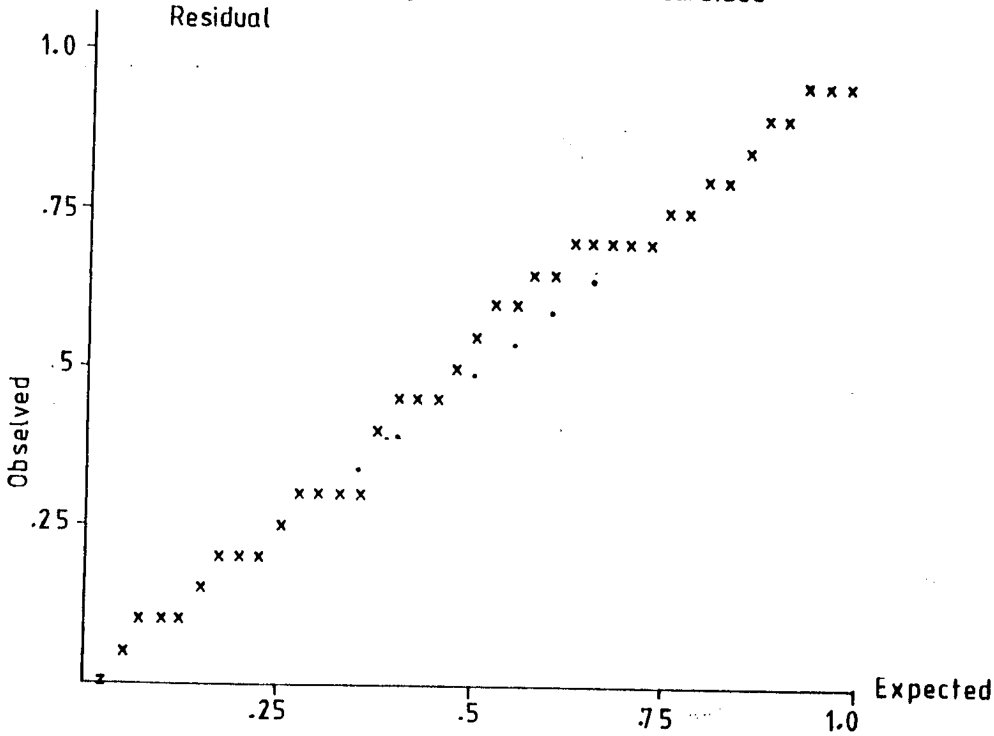
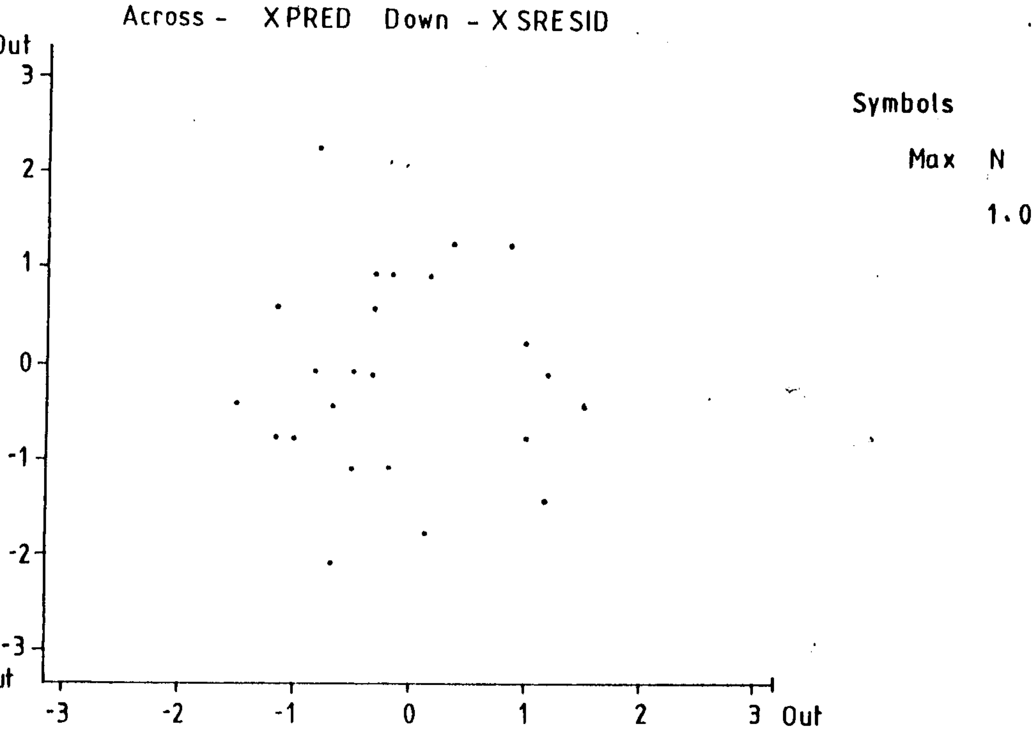


Fig. 4.5 (e)

Standardized Scatterplot

Across - XPRED Down - X SRESID



The Equation is :

$$\hat{Y}_2 = 25.28937 + 1.6345X_3 \qquad \text{Where } X_3 = 2.9 \text{ inches}$$

$$\hat{Y}_2 = 25.28937 + 1.6345 \times 2.9$$

$$\underline{\hat{Y}_2 = 30.04}$$

That means, given the total rainfall up to 30th November 1985, the total rainfall for the whole rainy season would be 30.04 inches (763.04mm).

A 100 (1 - α) percent prediction interval for future observation given X_3 is:

$$\hat{Y}_2 \pm t_{\alpha/2},\ _{k-n}\sqrt{MS_E\left[1 + \frac{1}{n} + \frac{(X_o - \bar{X})^2}{S_{xx}}\right]} \ , \quad \text{where } X_o = X_3$$

$$30.04 \pm 2.069\sqrt{31.18216\left[1 + \frac{1}{25} + \frac{(2.9 - 3.81)^2}{90.8}\right]}$$

$$\underline{30.04 \pm 11.83}$$

That means, we are 95% confident that given the total rainfall up to 30th November, 1985 the total rainfall for 1985/86 rainy season would be between 18.21 inches (462.53mm) and 41.87 inches (1063.50mm).

4.5.1.3 <u>Predicting Y_2 from X_4</u>

The scatter plot on Fig 4.5f shows that the total rainfall of the rainy season is significantly related to the total rainfall up to 31st December. Table 4.5e shows the Analysis of variance for Y_2 against X_4. From the ANOVA table, significance $F = 0.0328 < 0.05$. Figs 4.5(g)(h)(i) show the model adequacy using the probability plot and standardized scatter plot and the plot of standardized residuals. Table 4.5 (f) shows the t-values of each variable.

ANALYSIS OF VARIANCE

TABLE 4.5(e)

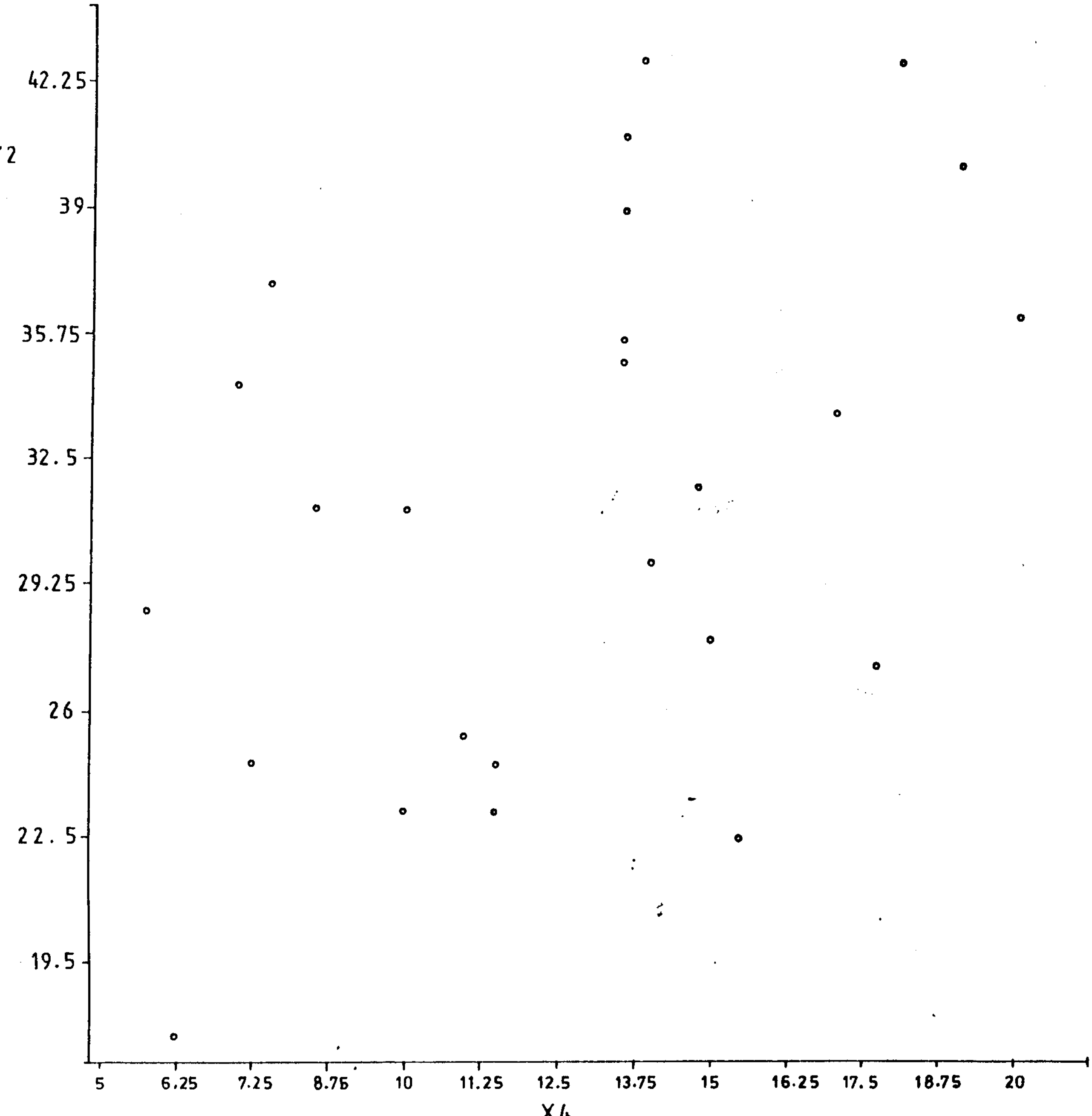
	DF	SUM OF SQUARES	MEAN SQUARE
Regression	1	214.15781	214.15781
Residuals	23	954.88052	41.51654
F = 5.15837		Significance F = 0.0328*	

TABLE 4.5(f)

Variable	β	S.E. β	95% Confidence interval of β	t-value	Sig t
X_4	0.73620	0.32415	0.06565-1.40675	2.271	0.0328

Fig 4.5(f)

PLOT OF Y2 WITH X4



25 Cases plotted

Fig: 4.5(g)
Normal Probability (P-P) Plot
Standardized Residual

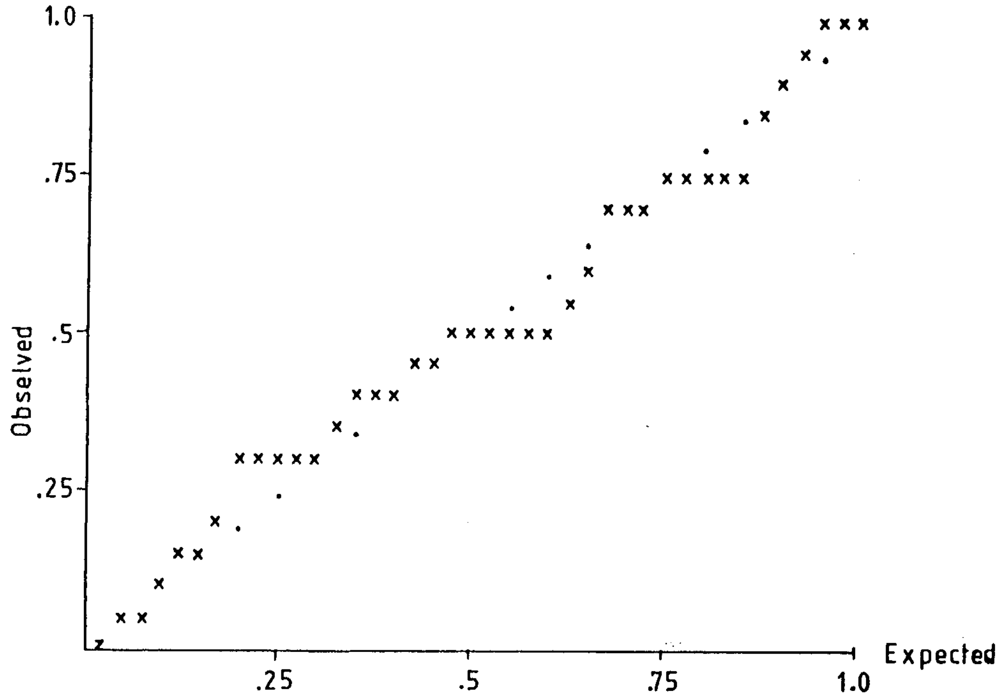


Fig 4.5 (h)
Standardized Scatterplot
Across - x PRED Down - x SRESID

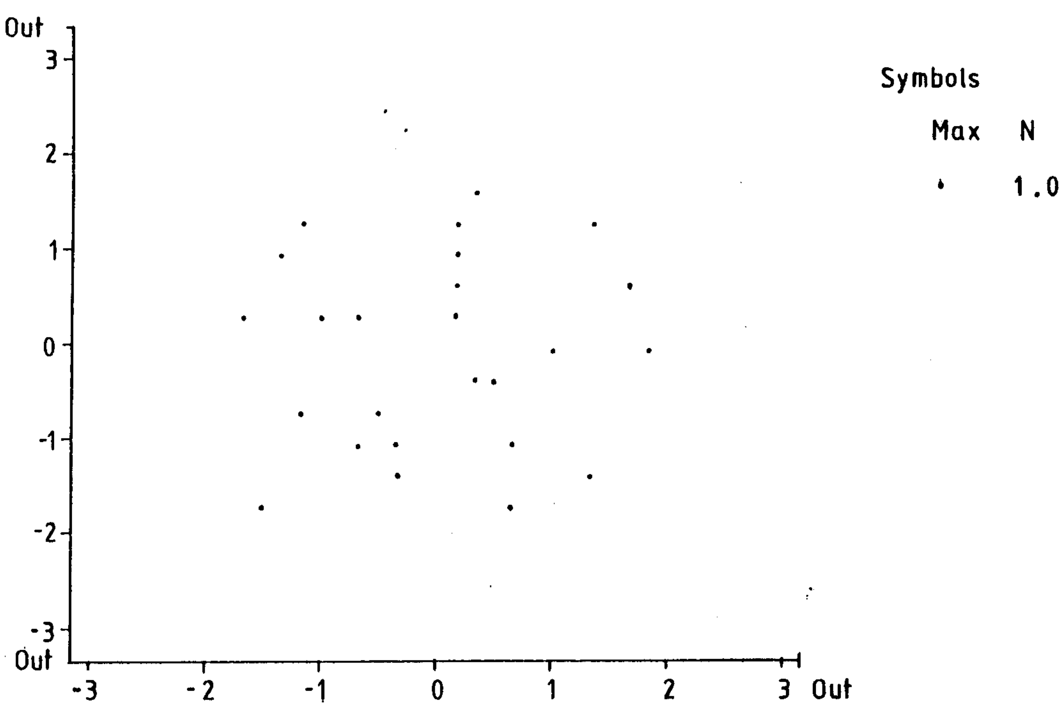


TABLE 4.5(i)

```
     * * * *  MULTIPLE  REGRESSION  * * * *

Equation Number 1     Dependent Variable .    Y2

Casewise Plot of Standardized Residual

*:  Selected M:  Missing

              - 3.0    0.0     3.0
 Case          O:.........:.........:O      Y2       *PRED       *RESID      *SEPRED

     1          .      *  .         .     27.67    33.3574     -5.6874      1.5173
     2          .         .  *      .     40.30    36.2949      4.0051      2.4590
     3          .         .   *     .     43.16    35.4629      7.6971      2.1556
     4          .      *  .         .     24.78    30.7513     -5.9713      1.3345
     5          .      *  .         .     23.50    29.6469     -6.1469      1.5343
     6          .         .   *     .     37.01    28.0052      9.0048      2.0201
     7          .         .*        .     31.28    29.5881      1.6919      1.5486
     8          .        *.         .     24.72    27.8064     -3.0864      2.0882
     9          .         *         .     36.53    37.0237     -0.4937      2.7374
    10          .   *     .         .     22.75    33.5562    -10.5562      1.5652
    11          .        *.         .     31.78    33.0629     -1.2829      1.4530
    12          .         .*        .     31.42    28.5058      2.9142      1.8557
    13          .    *    .         .     17.74    26.8420     -9.1020      2.4365
    14          .         . *       .     35.99    32.1721      3.8179      1.3185
    15          .         .  *      .     39.07    32.2457      6.8243      1.3258
    16          .         .    *    .     42.91    32.3341     10.5759      1.3354
    17          .        *.         .     29.60    32.4519     -2.8519      1.3500
    18          .         .  *      .     40.87    32.1206      8.7494      1.3139
    19          .    *    .         .     27.13    35.3010     -8.1710      2.0988
    20          .         *         .     34.00    34.7194      -.7194      -.9032
    21          .         .*        .     34.97    32.1132      2.8568      1.3133
    22          .         .*        .     28.62    26.5475      2.0725      2.5474
    23.         .         .  *      .     34.43    27.5046      6.9254      2.1943
    24          .    *    .         .     23.17    30.7660     -7.5960      1.3328
    25          .       * .         .     25.06    30.2801     -5.2201      1.4027
 Case          O:.........:.........:O      Y2       *PRED       *RESID      *SEPRED
              -3:0     0.0     3.0
```

The Equation is:

$$\hat{Y}_2 = 22.22603 + 0.73620X_4$$

$$\hat{Y}_2 = 22.22603 + 0.73620X7.7$$

$$\hat{Y}_2 = 27.89 \text{ inches}$$

That means, given total rainfall up to 31st December 1985, the total rainfall for 1985/86 rainy season would be 27.89 inches (708.53mm)

A 100 (1-α) percent prediction interval on future observation at X_4 is:

$$\hat{Y}_2 \pm t_{\alpha/2,\, n-2} \sqrt{MS_E \left[1 + \frac{1}{n} + \frac{(X_4 - X)^2}{S_{xx}} \right]}$$

$$27.89 + 2.069 \sqrt{41.51654 \left[1 + \frac{1}{25} + \frac{(7.7 - 12.65)^2}{395} \right]}$$

$$27.89 \pm 2.069 \sqrt{41.51654 \left[\frac{24}{25} + \frac{24.5028}{395} \right]}$$

$$\underline{27.89 \pm 14}$$

That means, we are 95% Confident that given the total rainfall up to 31st December 1985, the total rainfall for the rainy season 1985/86 will be between 13.89 inches (352.81mm) and 41.89 inches (1064.01mm)

TABLE 4.5.2

Dependant Variable Y_2					
	DF	SS	MS	F	Sig F
Regression	1	65.40586	65.40586	1.36308	0.2550
Residual	23	1103.63247	47.98402		
Non Significant					
X_1X_2					
Regression	2	228.84461	114.42231	2.67742	0.0910
Residual	22	940.19372	42.73608		
Non Significant					
X_1X_3					
Regression	2	243.02313	121.51156	2.88684	0.0770
Residual	22	946.01521	42.09160		
Non Significant					
X_1X_4					
Regression	2	222.95438	111.47719	2.59226	0.975
Residual	22	946.08396	43.00382		
Non Significant					
X_3X_4					
Regression	2	276.80151	138.40075	3.41257	0.0512
Residual	22	892.23683	40.55622		
Non Significant					
$X_1X_3X_4$					
Regression	3	277.00946	92.33649	2.17377	0.1213
Residual	21	892.02888	42.47757		

Non Significant

Table 4.5.2 shows the variables which are not significantly related to Y_2.

4.5.2 MULTIPLE LINEAR REGRESSION

Consider the model

$$\hat{Y}_2 = \hat{\beta}_o + \sum_{ij}^{4} \hat{\beta}_j X_i + \varepsilon, \ i = 1,2,3,4, \ j = 1,2,3,4$$

where X_i is a combination of the X_i's (observed values for 1985/86 rainy season)

4.2.1 Predicting Y_2 from X_2, X_3

The X matrix and Y vector are defined as follows:-

$$X = \begin{bmatrix} 1 & 46 & 6.17 \\ 1 & 38 & 8.97 \\ 1 & 47 & 5.54 \\ 1 & 34 & 3.38 \\ 1 & 38 & 2.98 \\ 1 & 46 & 3.5 \\ 1 & 48 & 1.4 \\ 1 & 40 & 2.01 \\ 1 & 50 & 5.64 \\ 1 & 21 & 4.20 \\ 1 & 45 & 5.90 \\ 1 & 57 & 3.13 \\ 1 & 32 & 2.62 \\ 1 & 36 & 3.30 \\ 1 & 60 & 4.03 \\ 1 & 36 & 2.15 \\ 1 & 26 & 2.96 \\ 1 & 45 & 4.33 \\ 1 & 47 & 2.48 \\ 1 & 41 & 6.63 \\ 1 & 32 & 6.12 \\ 1 & 32 & 2.20 \\ 1 & 51 & 3.19 \\ 1 & 24 & 1.50 \\ 1 & 49 & 1.02 \end{bmatrix} \qquad Y_2 = \begin{bmatrix} 27.67 \\ 40.30 \\ 43.16 \\ 24.78 \\ 23.50 \\ 37.01 \\ 31.28 \\ 24.72 \\ 36.53 \\ 22.73 \\ 31.78 \\ 31.42 \\ 17.74 \\ 35.99 \\ 39.07 \\ 42.91 \\ 29.60 \\ 40.87 \\ 27.13 \\ 34.00 \\ 34.97 \\ 28.62 \\ 34.43 \\ 23.17 \\ 25.06 \end{bmatrix}$$

$$X'X = \begin{bmatrix} 1 & \ldots & 1 \\ 46 & \ldots & 49 \\ & & \\ .17 & \ldots & 1.02 \end{bmatrix} \begin{bmatrix} 1 & 46 & 6.17 \\ . & . & . \\ . & . & . \\ . & . & . \\ 1 & 49 & 1.02 \end{bmatrix} \begin{bmatrix} 25 & 1021 & 95.35 \\ 1021 & 44097 & 3932.88 \\ & & \\ 95.35 & 3932.88 & 484.23 \end{bmatrix}$$

$$Adj(X'X)^T = \begin{bmatrix} 5{,}885{,}545.22 & -119{,}398 & -189{,}178.47 \\ -119{,}398.73 & 3014.13 & -969.65 \\ -189{,}178.47 & -969.65 & 59{,}984 \end{bmatrix}$$

$\underline{X'X = 7{,}194{,}360.06}$

$$(X'X)^{-1} = \frac{Adj\ (X'X)^T}{X'X} = \begin{bmatrix} 0.818077 & -0.016596 & -0.026295 \\ -0.016596 & 0.000418 & -0.000134 \\ -0.026295 & -0.000134 & 0.008337 \end{bmatrix}$$

TABLE 4.5.2(a)

ANALYSIS OF VARIANCE

	DF	SUM OF SQUARES	MEAN SQUARES
Regression	2	411.28467	205.64233
Residual	22	757.75367	34.44335
F = 5.97045		Significance F = 0.0085**	

TABLE 4.5.2(b)

Variable	β	S.E. β	95% Confidence interval of β	t-value	Sig t
X_3	1.52233	0.62028	0.23594 - 2.80872	2.454	0.0225*
X_2	0.26993	0.12178	0.01739 - 0.52248	2.217	0.0373*

Fig 4.5.2.1 (a)

Normal Probability (P-P) Plot

Standardized Residual

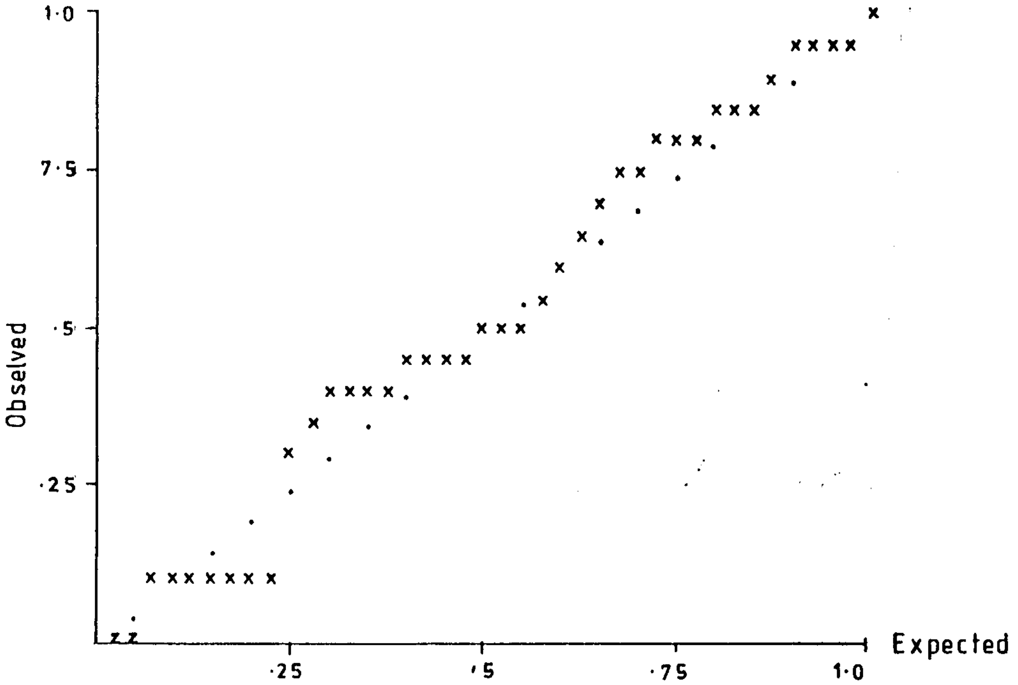


Fig 4.5.2.1 (b)

Standardized Scatterplot

Across XPRED Down - XSRESID

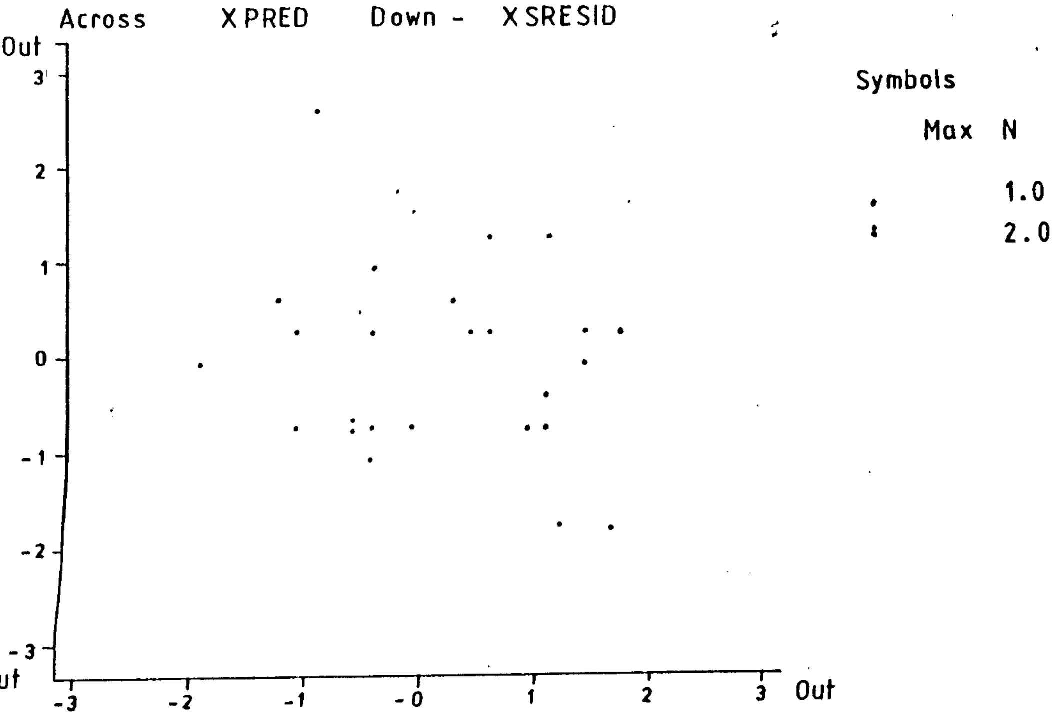


FIG. 4.5.2.1(c)

```
**** MULTIPLE REGRESSION ****

Equation Number 1,          Dependent Variable.  Y2

Casewise Plot of Standardized Residual

*: Selected      M; Missing

           -3.0       0.0       3.0
Case        0:.........:.........:0       Y2      *PRED     *RESID    *SERRED
      1     .     *    .         .     27.67    36.5179    -8.8479     1.9373
      2     .          .*        .     40.30    38.6209     1.6791     3.4514
      3     .          .   *     .     43.16    35.8287     7.3313     1.7178
      4     .        * .         .     24.78    29.0314    -4.2514     1.4512
      5     .       *  .         .     23.50    29.5022    -6.0022     1.3171
      6     .          .  *      .     37.01    32.4532     4.5568     1.3532
      7     .          .*        .     31.28    29.7962     1.4838     2.1449
      8     .        * .         .     24.72    28.5654    -3.8454     1.6190
      9     .          *         .     36.53    36.7908     -.2608     1.9213
     10     .        * .         .     22.75    26.7706    -4.0206     2.7148
     11     .        * .         .     31.78    35.8369    -4.0569     1.7883
     12     .        * .         .     31.42    34.8592    -3.4392     2.3604
     13     .     *    .         .     17.74    27.3345    -9.5945     1.7177
     14     .          .   *     .     35.99    29.4495     6.5405     1.3398
     15     .          .*        .     39.07    37.0391     2.0309     2.6051
     16     .          .        *.     42.91    27.6988    15.2112     1.6394
     17     .          . *       .     29.60    26.2325     3.3675     2.1823
     18     .          .   *     .     40.87    33.4468     7.4232     1.3074
     19     .        * .         .     27.13    31.1704    -4.0404     1.6523
     20     .         *.         .     34.00    35.8685    -1.8685     2.1032
     21     .          .*        .     34.97    32.6627     2.3073     2.2006
     22     .          .*        .     28.62    26.6952     1.9248     1.8322
     23     .          .*        .     34.43    33.3310     1.0990     1.7718
     24     .          *         .     23.17    23.4701     -.3001     2.6732
     25     .       *  .         .     25.06    29.4877    -4.4277     2.3790
Case        0:.........:.........:0       Y2      *PRED     *RESID    *SERRED
           -3.0       0.0       3.0
```

Table 4.5.2(a) shows the ANOVA table with significance F $= 0.0085 < 0.01$ that is, the toal rainfall is highly significantly related to X_2, X_3. Table 4.5.2(b) shows the t - values of each variable. Figs 4.5.2.1(a), 4.5.2(b)(c) shows the model adequacy using the normal probability plot, standardized scatter plot, and standardized residuals respectively.

$$\begin{bmatrix} \beta_u \\ \hat{\beta}_2 \\ \hat{\beta}_3 \end{bmatrix} = \begin{bmatrix} 14.70821 \\ 0.25993 \\ 1.52233 \end{bmatrix}$$

The least squares fit (with the regression coefficients reported to five decimals) is:

$$\hat{Y}_2 = 14.70821 + 0.26993X_2 + 1.52233X_3$$

$$\hat{Y}_2 = 14.70821 + 0.26993X37 + 1.52233X2.9$$

$$\underline{\hat{Y}_2 = 29.11}$$

That is, the predicted total rainfall for 1985/86 rainy season given X_2 and X_3 is 29.11 inches (739.40mm).

A 100 $(1 - \alpha)$ per cent prediction interval for this future observation is

$$\hat{Y}_2 \pm t_{\alpha/2},\, k-n \sqrt{MS_E \left[1 + X_o^1 \, (X^1X)^{-1} X_o \right]}$$

$$X_o' \, (X X^{-1}) = 1 \quad 37 \quad 2.9 \begin{bmatrix} 0.818077 & -0.016596 & 0.026295 \\ -0.016596 & 0.000418 & -0.000134 \\ -0.026295 & -0.000134 & 0.000833 \end{bmatrix} \begin{bmatrix} 1 \\ 37 \\ 2.9 \end{bmatrix}$$

$X_o'(X X)^{-1}X_o \qquad = 0.0510618$

$$Y_2 \pm t_{\alpha/2}\ , \ k-n \sqrt{MS_E \left[1 + X_o^1 \ (X^1X)^{-1}X_o \right]}$$

$$29.11 \pm t_{0.025,\ 22} \sqrt{34.44335 \left[1 + 0.0510618 \right]}$$

$$29.11 \pm 2.074 \sqrt{36.202089}$$

$$\underline{29.11 + 12.48}$$

That means that predicted interval for the total rainfall for 1985/86 rainy season given X_2, X_3 is between 16.63 inches (422.40mm) and 41.59 inches (1056.39mm).

4.5.2.2 Predicting Y_2 from X_2, X_4

The X-matrix and Y vector are defined as follows:

$$X = \begin{bmatrix} 1 & 46 & 15.12 \\ 1 & 38 & 19.11 \\ 1 & 47 & 17.98 \\ 1 & 34 & 11.58 \\ 1 & 38 & 10.08 \\ 1 & 46 & 7.85 \\ 1 & 48 & 10.00 \\ 1 & 40 & 7.58 \\ 1 & 50 & 20.10 \\ 1 & 21 & 15.39 \\ 1 & 45 & 14.72 \\ 1 & 57 & 8.53 \\ 1 & 32 & 6.27 \\ 1 & 36 & 13.51 \\ 1 & 60 & 13.61 \\ 1 & 36 & 13.73 \\ 1 & 26 & 13.89 \\ 1 & 45 & 13.44 \\ 1 & 47 & 17.76 \\ 1 & 41 & 16.97 \\ 1 & 32 & 13.43 \\ 1 & 32 & 5.87 \\ 1 & 51 & 7.17 \\ 1 & 24 & 11.60 \\ 1 & 49 & 10.94 \end{bmatrix} \qquad Y_2 = \begin{bmatrix} 27.67 \\ 40.30 \\ 43.16 \\ 24.78 \\ 23.50 \\ 37.01 \\ 31.28 \\ 24.72 \\ 36.53 \\ 22.75 \\ 31.78 \\ 31.42 \\ 17.14 \\ 35.99 \\ 39.07 \\ 42.91 \\ 29.60 \\ 40.87 \\ 27.13 \\ 34.00 \\ 34.97 \\ 28.62 \\ 24.43 \\ 23.17 \\ 25.06 \end{bmatrix}$$

$$X'X = \begin{bmatrix} 1 & \cdots & 1 \\ 46 & \cdots & 49 \\ & & \\ 15.12 & \cdots & 10.94 \end{bmatrix} \begin{bmatrix} 1 & 46 & 15.6 \\ . & . & . \\ & & \\ 1 & 49 & 10.94 \end{bmatrix} = \begin{bmatrix} 25 & 1021 & 316.23 \\ 1021 & 44097 & 12958.94 \\ & & \\ 316.23 & 12958.94 & 4394.97 \end{bmatrix}$$

$$Adj(X'X)^T = \begin{bmatrix} 25{,}870{,}866.16 & -389{,}258.77 & -713{,}716.57 \\ -389{,}258.77 & 9{,}872.83 & -1{,}102.67 \\ -713{,}716.57 & -1{,}102.67 & 59{,}984 \end{bmatrix}$$

$$\underline{|X'X| \;=\; 818{,}506{,}267.23}$$

$$(X'X)^{-1} = \frac{Adj(X'X)^T}{X'X} = \begin{bmatrix} 0.031607 & -0.000475 & -0.000871 \\ -0.000475 & 0.000012 & -0.000001 \\ -0.000871 & -0.000001 & 0.000073 \end{bmatrix}$$

TABLE 4.5.2(c)

ANALYSIS OF VARIANCE

	DF	SUM OF SQUARES	MEAN SQUARE
Regression	2	400.55777	200.27889
Residual	22	768.48056	34.93093
F = 5.73357		Significance F = 0.0099**	

TABLE 4.5.2(d)

Variable	β	S.E.β	95% Confidence interval of β	t-value	Sig t
X_4	0.70629	0.29761	0.08909-1.32350	2.273	0.0268*
X_2	0.28254	0.12231	0.02888-0.53620	2.310	0.0307*

Fig 4.5.2.2 (b)

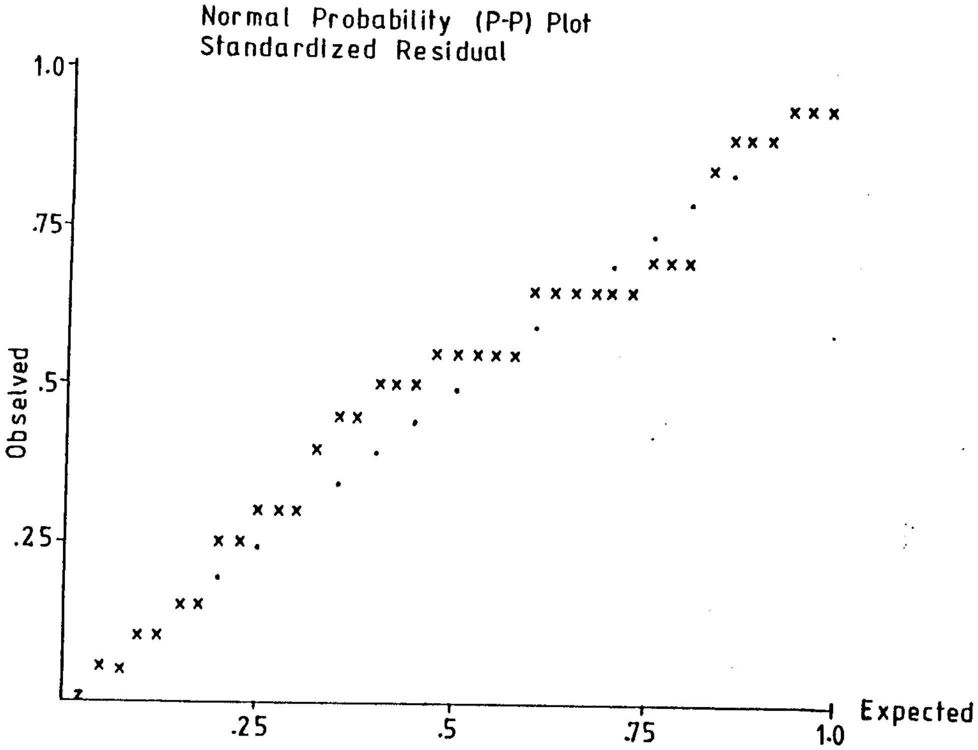


Fig 4.5.2.2 (b)

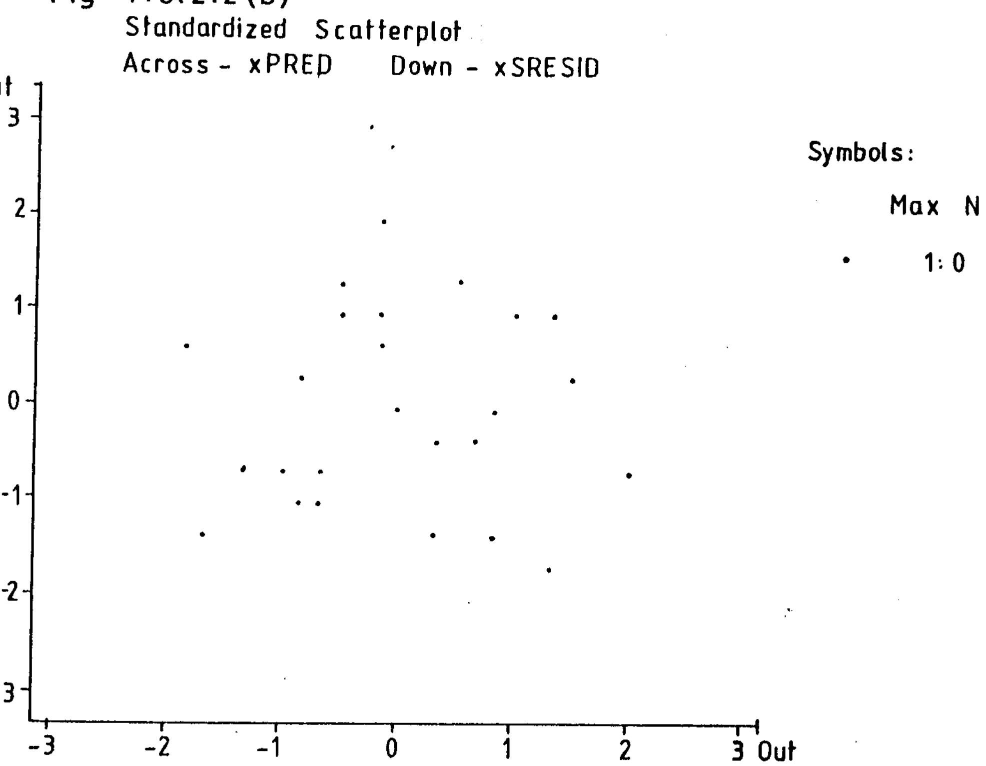


Table 4.5.2(c) shows the ANOVA table with significance $F = 0.0099 < 0.01$ which means that Y_2 is highly significantly related to X_2, X_4. Table 4.5.2(d) shows the t-values of each variable. Figs 4.5.2.2(b), 4.5.2.2(c) shows the model adequacy using the Normal probability plot and standardized scatter plot.

$$\begin{bmatrix} \hat{\beta}_0 \\ \hat{\beta}_2 \\ \hat{\beta}_4 \end{bmatrix} = \begin{bmatrix} 11.06527 \\ 0.28254 \\ 0.70629 \end{bmatrix}$$

Therefore, the least square fit is:

$$\hat{Y}_2 = 11.06527 + 0.28254X_2 + 0.70629X_4$$

$$\hat{Y}_2 = 11.06527 + 0.28254X37 + 0.70629X7.68$$

$$\underline{\hat{Y}_2 = 26.93}$$

That means, the predicted total rainfall for 1985/86 rainy season given X_2, X_4 is 26.93 inches (684.02mm)

A 100 $(1 - \alpha)$ per cent prediction interval for the future Observation is

$$\hat{Y}_2 \pm t_{\alpha/2}, k-n \sqrt{MS_E \left[1 + X_o^1 (X^1X)^{-1}X_o\right]}$$

4.4.2.3 Predicting Y_2 from X_1, X_2, X_3

The X matrix and Y_2 vector are defined as follows:-

$$X = \begin{bmatrix} 1 & 52 & 46 & 6.17 \\ 1 & 46 & 38 & 8.97 \\ 1 & 52 & 47 & 5.54 \\ 1 & 51 & 34 & 3.38 \\ 1 & 53 & 38 & 2.98 \\ 1 & 58 & 46 & 3.50 \\ 1 & 74 & 48 & 1.40 \\ 1 & 85 & 40 & 2.01 \\ 1 & 65 & 50 & 5.64 \\ 1 & 28 & 21 & 4.20 \\ 1 & 40 & 45 & 5.90 \\ 1 & 46 & 57 & 3.13 \\ 1 & 78 & 32 & 2.62 \\ 1 & 78 & 36 & 3.30 \\ 1 & 52 & 60 & 4.03 \\ 1 & 58 & 36 & 2.15 \\ 1 & 68 & 26 & 2.96 \\ 1 & 54 & 45 & 4.33 \\ 1 & 61 & 47 & 2.48 \\ 1 & 65 & 41 & 6.63 \\ 1 & 69 & 32 & 6.12 \\ 1 & 91 & 32 & 2.20 \\ 1 & 43 & 51 & 3.19 \\ 1 & 79 & 24 & 1.50 \\ 1 & 71 & 49 & 1.02 \end{bmatrix} \qquad Y_2 = \begin{bmatrix} 27.67 \\ 40.30 \\ 43.16 \\ 24.78 \\ 23.50 \\ 37.01 \\ 31.28 \\ 24.72 \\ 36.53 \\ 22.75 \\ 31.78 \\ 31.42 \\ 17.74 \\ 35.99 \\ 39.07 \\ 42.91 \\ 29.60 \\ 40.87 \\ 27.13 \\ 34.00 \\ 34.97 \\ 28.62 \\ 34.43 \\ 23.17 \\ 25.06 \end{bmatrix}$$

$$X'X = \begin{bmatrix} 1 & \ldots & 1 \\ 52 & \ldots & 71 \\ 46 & \ldots & 49 \\ & & \\ 6.17 & & 1.02 \end{bmatrix} \begin{bmatrix} 1 & 52 & 46 & 6.17 \\ . & . & . & . \\ . & . & . & . \\ . & . & . & . \\ 1 & 71 & 49 & 1.02 \end{bmatrix}$$

$$X'X = \begin{bmatrix} 25 & 1517 & 1021 & 95.35 \\ 1517 & 105183 & 61142 & 5457.41 \\ 1021 & 61142 & 44097 & 102958.94 \\ 95.35 & 5457.41 & 3932.88 & 484.23 \end{bmatrix}$$

$$(X'X)^{-1} = \begin{bmatrix} 12.52500007 & 0.2069942 & -0.27552352 & -0.00616346 \\ 0.02069912 & 0.00014242 & -0.00130134 & -0.00003035 \\ -0.27552352 & -0.00130134 & -0.00006093 & -0.00000973 \\ -0.00616346 & -0.00003035 & -0.00000973 & -0.00123740 \end{bmatrix}$$

$$X_o'(X'X)^{-1}X_o = \begin{bmatrix} 1 & 41 & 37 & 2.9 \end{bmatrix} \begin{bmatrix} 12.52500 & 0.02070 & -0.27552 & -0.00616 \\ 0.02020 & 0.00014 & -0.00130 & -0.00003 \\ -0.27552 & -0.00130 & -0.00006 & -0.00001 \\ -0.00616 & -0.00003 & -0.00001 & -0.00124 \end{bmatrix} \begin{bmatrix} 1 \\ 41 \\ 37 \\ 2.9 \end{bmatrix}$$

$$\underline{X_o'(X'X)^{-1}\, X_o = 0.051297}$$

TABLE 4.5.2.3(a)

ANALYSIS OF VARIANCE

	DF	SUM OF SQUARE	MEAN SQUARE
Regression	3	414.49442	138.16481
Residual	21	754.54391	35.93066
F = 3.84532		Significance F = 0.0244	

TABLE 4.5.2.3(b)

Variance	β	S.E. β	95% Confidence Interval of β	t-value	Sig.t
X_3	1.61985	0.71263	0.13787 – 3.10184	2.273	0.0336
X_2	0.27794	0.12723	0.01335 – 0.54253	2.185	0.0404
X_1	0.02772	0.09274	-0.16515 – 0.22059	0.299	0.7680

Fig 4·5·2·3b

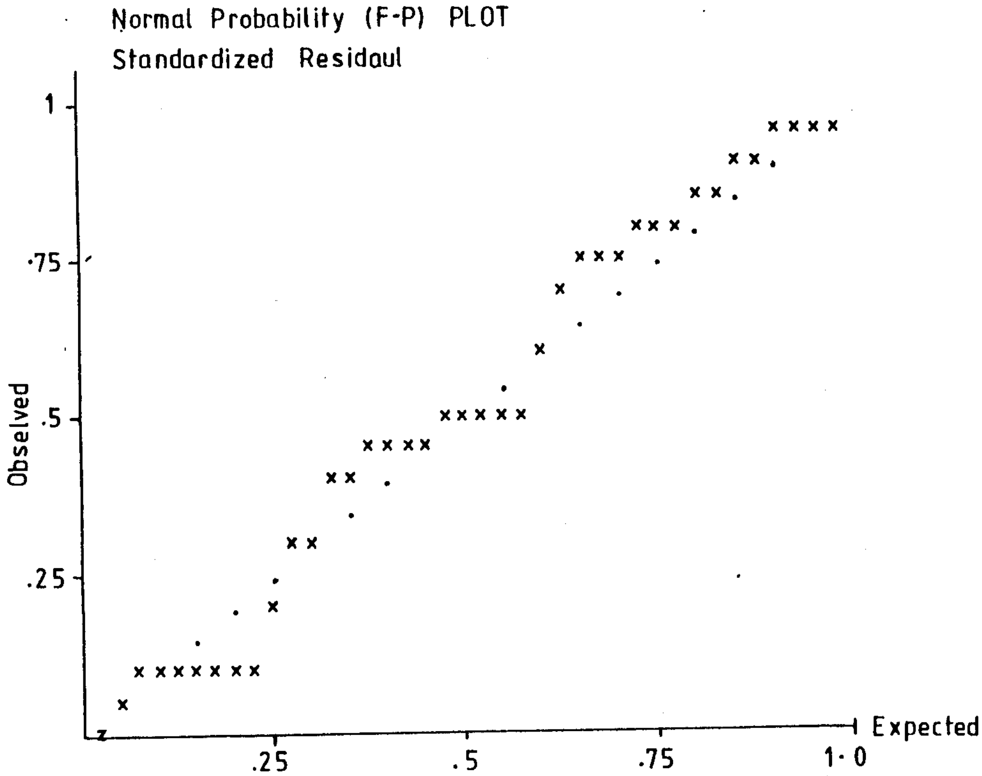


Fig 4·5·2·3c

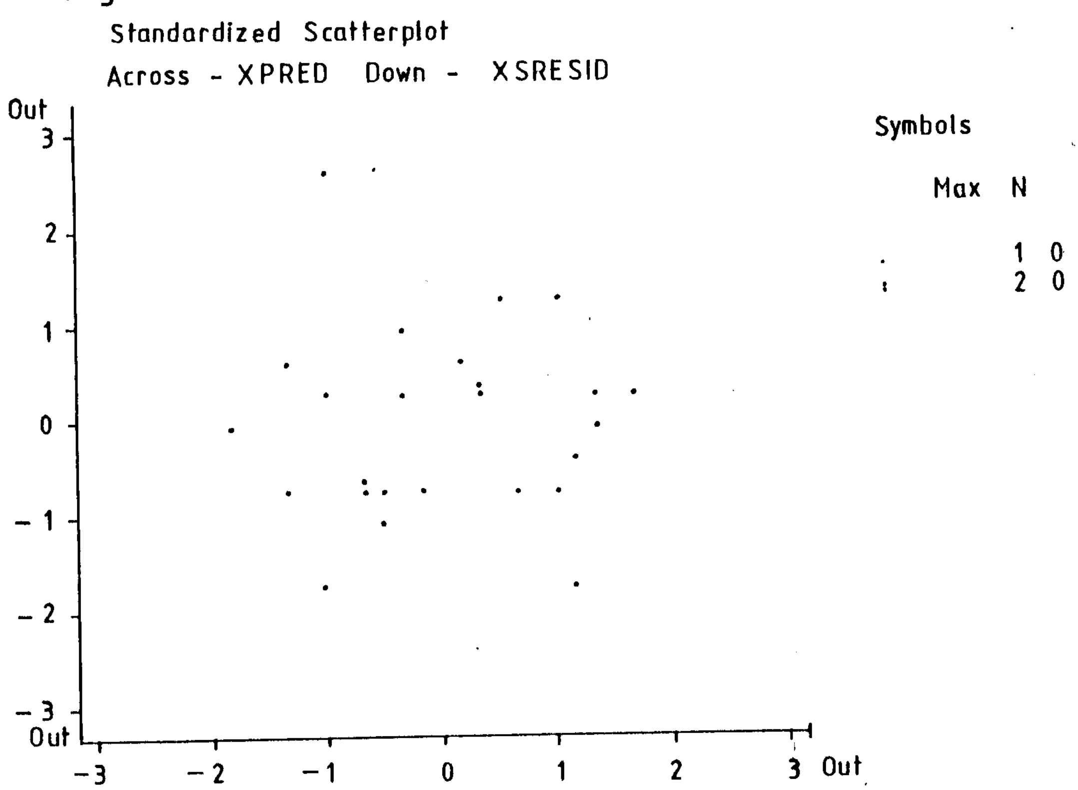


Table 4.4.2.3(a) shows the ANOVA table with significance F 0.0244 0.05, which means Y_2 is significantly related by X_1, X_2, X_3, although the t value for the variable X_1 the P value is greater than 0.05.

Figs 4.4.2.3(b), 4.4.2.3(c) shows the model adequacy using the Normal probability plot and standardized scatter plot.

$$\begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \beta_3 \end{bmatrix} = \begin{bmatrix} 12.32716 \\ 0.02772 \\ 0.27794 \\ 1.61985 \end{bmatrix}$$

The least square fit is:

$$\hat{Y}_2 = 12.32716 + 0.02772X_1 + 0.27794X_2 + 1.161985X_3$$

$$\hat{Y}_2 = 12.32716 + 0.02772x41 + 0.27794x37 + 1.161985x2.9$$

$$\hat{Y}_2 = 27.12$$

The predicted total rainfall for 1985/86 rainy season is 27.12 inches (688.85mm).

A 100 (1-α) percent predicted interval for 1985/86 total rainfall is:

$$Y_2 \pm t_{\alpha/2}\,,\, k\text{-}n \sqrt{MS_E\left[1 + X_o^{'}\,(X^{'}X)^{-1}X_o\right]}$$

$$27.12 \pm t_{0.025,21}\sqrt{35.93066\ (1 + 0.051297)}$$

$$27.12 \pm 12.78$$

The predicted interval for the total rainfall for 1985/86 rainy season given X_1, X_2, X_3 is between 14.34 inches (364.24mm) and 39.9 inches (1013.46mm)

4.4.2.4 Predicting Y_2 from $X_2, X_3 X_4$

The X matrix and Y_2 vector are defined as:

$$X = \begin{bmatrix} 1 & 46 & 6.17 & 15.12 \\ 1 & 38 & 8.97 & 19.11 \\ 1 & 47 & 5.54 & 17.98 \\ 1 & 34 & 3.38 & 11.58 \\ 1 & 38 & 2.98 & 10.08 \\ 1 & 46 & 3.50 & 7.85 \\ 1 & 48 & 1.40 & 10.00 \\ 1 & 40 & 2.01 & 7.58 \\ 1 & 50 & 5.64 & 20.10 \\ 1 & 21 & 4.20 & 15.39 \\ 1 & 45 & 5.90 & 14.72 \\ 1 & 57 & 3.13 & 8.53 \\ 1 & 32 & 2.62 & 6.27 \\ 1 & 36 & 3.30 & 13.51 \\ 1 & 60 & 4.03 & 13.61 \\ 1 & 36 & 2.15 & 13.73 \\ 1 & 26 & 2.96 & 13.89 \\ 1 & 45 & 4.33 & 13.44 \\ 1 & 47 & 2.48 & 17.76 \\ 1 & 41 & 6.63 & 16.97 \\ 1 & 42 & 6.12 & 13.43 \\ 1 & 32 & 2.20 & 5.87 \\ 1 & 51 & 3.19 & 7.17 \\ 1 & 24 & 1.50 & 11.60 \\ 1 & 49 & 1.02 & 10.94 \end{bmatrix} \qquad Y_2 = \begin{bmatrix} 27.67 \\ 40.30 \\ 43.16 \\ 24.78 \\ 23.50 \\ 37.01 \\ 31.28 \\ 24.72 \\ 36.53 \\ 22.75 \\ 31.78 \\ 31.42 \\ 17.74 \\ 35.99 \\ 39.07 \\ 42.91 \\ 29.60 \\ 40.87 \\ 27.13 \\ 34.00 \\ 34.97 \\ 28.62 \\ 34.43 \\ 23.17 \\ 25.06 \end{bmatrix}$$

$$X'X = \begin{bmatrix} 1 \ldots 1 \\ 46 \ldots 49 \\ 6.17 \ldots 1.02 \\ \\ 15.12 \ldots 10.94 \end{bmatrix} \begin{bmatrix} 1 & 46 & 6.17 & 15.12 \\ . & . & . & . \\ . & . & . & . \\ . & . & . & . \\ 1 & 49 & 10.2 & 10.94 \end{bmatrix}$$

$$X'X = \begin{bmatrix} 25 & 1021 & 95.35 & 316.23 \\ 1021 & 44097 & 3932.88 & 102958.94 \\ 95.35 & 3932.88 & 484.23 & 1329.54 \\ 316.23 & 102958.94 & 1329.54 & 4394.97 \end{bmatrix}$$

$$(X'X)^{-1} = \begin{bmatrix} 0.0194700 & -0.00026400 & 0.00040000 & -0.00000600 \\ -0.0002640 & 0.00000001 & -0.00000015 & -0.00000014 \\ 0.0004000 & -0.00000015 & 0.0 & -0.00000014 \\ -0.0000060 & -0.00000014 & -0.00000014 & 0.0 \end{bmatrix}$$

$$\underline{X_o'(X'X)^{-1}X_o = 0.002158}$$

TABLE 4.4.2.4(a)

ANALYSIS OF VARIANCE

	DF	SUM OF SQUARES	MEAN SQUARE
Regression	3	448.47475	149.49158
Residual	21	720.56359	34.31255
F = 4.35676		Significance F = 0.0155	

TABLE 4.4.2.4(b)

Variable	β	S.E. β	95% Confidence interval of β	t-value	Sig. t
X_4	0.40557	0.38957	-0.405 – 1.216	1.041	0.3097
X_2	0.27190	0.12156	0.019 – 0.525	2.237	0.0363
X_3	0.96626	0.81767	-0.734 – 2.667	1.182	0.2505

Table 4.4.2.4(a) shows the ANOVA table with significance F = 0.0155

less than 0.05, although the t - values of the variable X_3 and X_4 p - values are greater than 0.05. Figs 4.5.2.4(c) and 4.4.3.4(d) shows the model adequacy using the Normal probability plot and standardized scatter plot respectively. Table 4.5.3.4(b) shows the t - values, and from this table we observe that X_2 is the only variable which contributes to the relationship of the variables to Y_2.

$$\begin{bmatrix} \beta_0 \\ \beta_2 \\ \beta_3 \\ \beta_4 \end{bmatrix} = \begin{bmatrix} 11.61853 \\ 0.27190 \\ 0.96626 \\ 0.40557 \end{bmatrix}$$

The least square fit is:

$$\hat{Y}_2 = 11.61853 + 0.27190X_2 + 0.96626X_3 + 0.40557X_4$$

$$\hat{Y}_2 = 11.61853 + 0.27190\text{x } 37 + 0.96626\text{x } 2.9 + 0.40557\text{x } 7.68$$

is 27.60 inches (700.93mm)

A 100(1 $-\alpha$) percent predicted interval for the total rainfall for 1985/86 rainy season is

$$\hat{Y}_2 \pm t_{\alpha/2}\text{, k-n} \sqrt{MS_E \left[1 + X_o'(X'X)^{-1}X_o \right]}$$

$$\hat{Y}_2 \pm t_{o.025,21} \sqrt{34.31255 \left(1 + 0.002158 \right)}$$

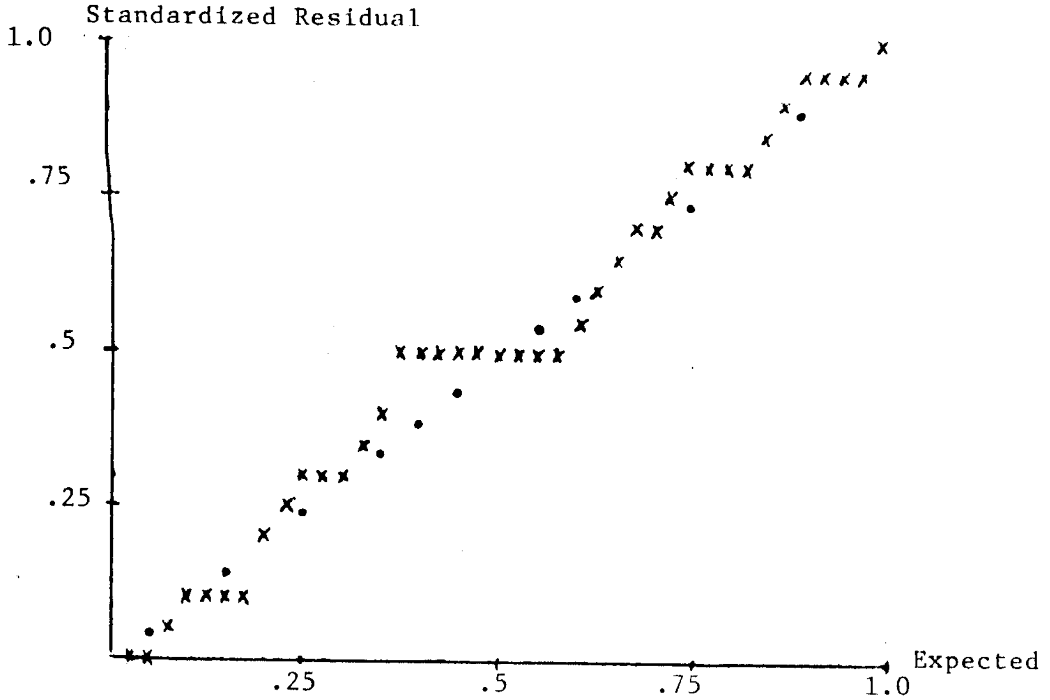
$$\underline{27.60 \pm 12.20}$$

The predicted interval for the total rainfall for 1985/96 rainy season given X_2, X_3, X_4 is between 15.16 inches (391.16mm) and 39.8 inches (1010.92mm).

fig 4.5.2.4c

Normal probability (P-P) Plot

Standardized Residual

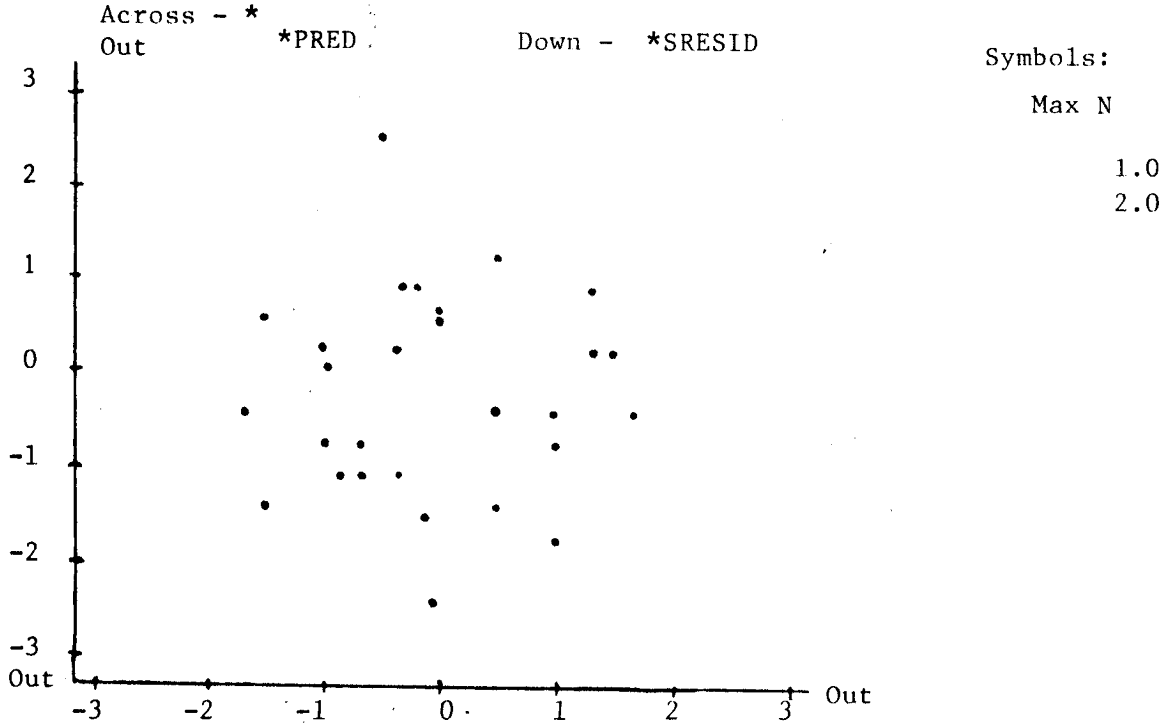


****

fig 4.5.2.4d

Standardized Scatterplot

Across - *PRED Down - *SRESID



4.4.2.5 Predicting Y_2 from X_1, X_2, X_3, X_4.

The X matrix and Y_2 vector are defined as follows

$$X = \begin{bmatrix} 1 & 52 & 46 & 6.17 & 15.12 \\ 1 & 46 & 38 & 8.97 & 19.11 \\ 1 & 52 & 47 & 5.54 & 17.98 \\ 1 & 51 & 34 & 3.38 & 11.58 \\ 1 & 53 & 38 & 2.98 & 10.08 \\ 1 & 58 & 46 & 3.50 & 7.85 \\ 1 & 74 & 48 & 1.40 & 10.00 \\ 1 & 85 & 40 & 2.01 & 7.58 \\ 1 & 65 & 50 & 5.64 & 20.00 \\ 1 & 28 & 21 & 4.20 & 15.39 \\ 1 & 40 & 45 & 5.90 & 14.72 \\ 1 & 46 & 57 & 3.13 & 8.53 \\ 1 & 78 & 32 & 2.62 & 6.27 \\ 1 & 78 & 36 & 3.30 & 13.51 \\ 1 & 52 & 60 & 4.03 & 13.61 \\ 1 & 58 & 36 & 2.15 & 13.73 \\ 1 & 68 & 26 & 2.96 & 13.89 \\ 1 & 54 & 45 & 4.33 & 13.44 \\ 1 & 61 & 47 & 2.48 & 17.76 \\ 1 & 65 & 41 & 6.63 & 16.97 \\ 1 & 69 & 32 & 6.12 & 13.43 \\ 1 & 91 & 32 & 2.20 & 5.87 \\ 1 & 43 & 51 & 3.19 & 7.17 \\ 1 & 79 & 24 & 1.50 & 11.60 \\ 1 & 71 & 49 & 1.02 & 10.94 \end{bmatrix} \qquad Y_2 = \begin{bmatrix} 27.67 \\ 40.30 \\ 43.16 \\ 24.78 \\ 23.50 \\ 37.01 \\ 31.28 \\ 24.72 \\ 36.53 \\ 22.75 \\ 31.78 \\ 31.42 \\ 17.74 \\ 35.99 \\ 39.07 \\ 42.91 \\ 29.60 \\ 40.87 \\ 27.13 \\ 34.00 \\ 34.97 \\ 28.62 \\ 34.43 \\ 23.17 \\ 25.06 \end{bmatrix}$$

$$X'X = \begin{bmatrix} 1 & \ldots & 1 \\ 52 & \ldots & 71 \\ 46 & \ldots & 49 \\ 6.17 & \ldots & 1.02 \\ 15.12 & \ldots & 10.94 \end{bmatrix} \begin{bmatrix} 1 & 52 & 46 & 6.17 & 15.12 \\ \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot \\ 1 & 71 & 49 & 1.02 & 10.94 \end{bmatrix}$$

$$X'X = \begin{bmatrix} 25 & 1517 & 1021 & 95.35 & 316.23 \\ 1517 & 105183 & 61142 & 5457.41 & 18649.16 \\ 1021 & 61142 & 44097 & 3932.88 & 102958.94 \\ 95.35 & 5457.41 & 3932.88 & 484.23 & 1329.54 \\ 316.23 & 18649.16 & 102958.94 & 1329.54 & 4394.97 \end{bmatrix}$$

$$\underline{X_o'(X'X)^{-1}X_o = 0.000419}$$

TABLE 4.4.2.5(a)

ANALYSIS OF VARIANCE

	DF	SUM OF SQUARES	MEAN SQUARE
Regression	4	454.17823	113.54456
Residual	20	714.86010	35.74301
F = 3.17669		Significance F = 0.0357	

TABLE 4.4.2.5(c)

Variable	β	S.E. β	95% confidence of β	t - Values	Sig. t
X_4	0.42089	0.39945	-0.412 - 1.254	1.054	0.3046
X_2	0.28270	0.12698	0.017 - 0.548	2.226	0.0376
X_3	0.03712	0.09293	-0.157 - 0.231	0.399	0.6938
X_3	1.07586	0.87848	-0.757 - 2.908	1.225	0.2349

Fig 4.5.2.5c

Normal Probability (P-P) Plot
Standardized Residual

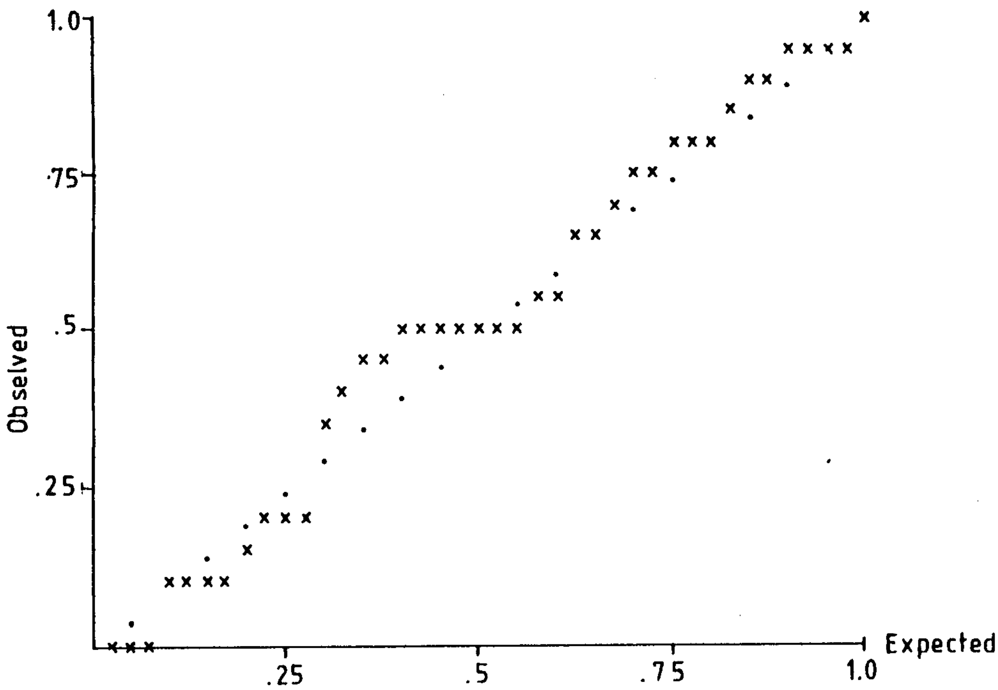


Fig 4.5.2.5d

Standardized Scatter plot
Across x PRED Down - x SRESID

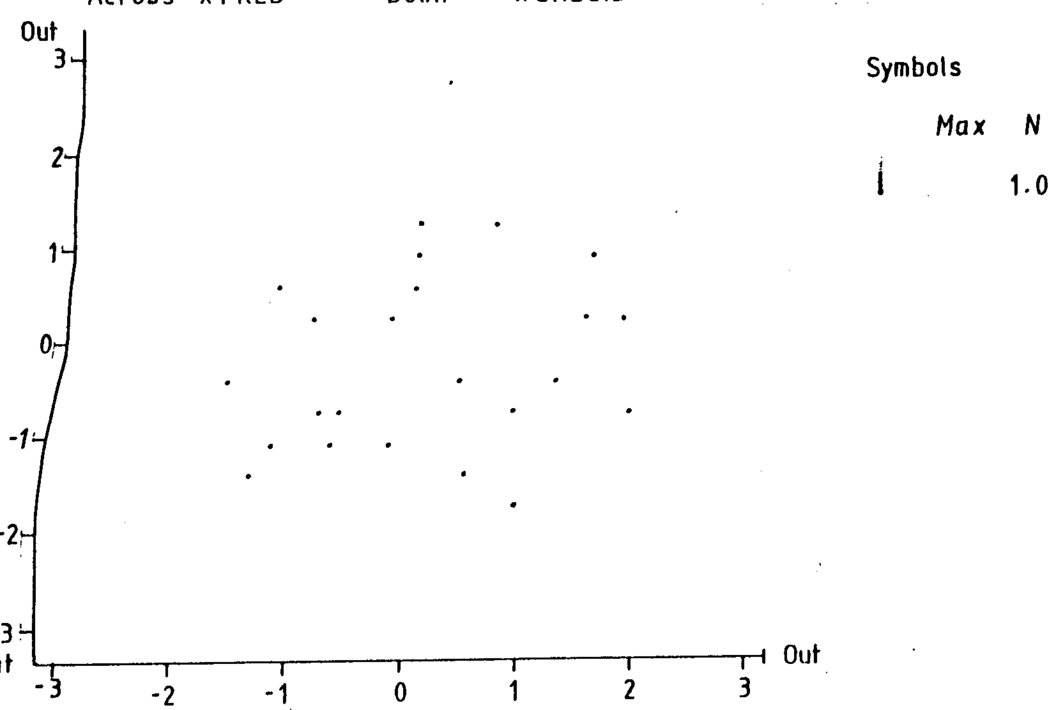


FIG. 4.5.2.5(e)

```
****  MULTIPLE  REGRESSION  ****

Equation Number 1     Dependent Variable  Y2

Casewise Plot of Standardized Residual
*:  Selected M: Missing

         -3.0      0.0       3.0
Case #    O:.........:.........:O        Y2      *PRED      *RESID    *SEPRED
    1     .      *   .         .      27.67    35.1055     -7.4355     2.1030
    2     .          *         .      40.30    39.7252       .5748     3.7996
    3     .          .    *    .      43.16    35.5382      7.6218     2.1844
    4     .      *   .         .      24.78    30.7204     -5.9404     1.7649
    5     .      *   .         .      23.50    29.6905     -6.1905     1.9125
    6     .          .    *    .      37.01    29.3496      7.6604     2.3738
    7     .          . *       .      31.28    27.8070      3.4730     2.1794
    8     .        * .         .      24.72    27.4478     -2.7278     2.1794
    9     .          *         .      36.53    36.3815       .1485     3.0847
   10     .    *     .         .      22.75    33.2486    -10.4986     3.4263
   11     .        * .         .      31.78    34.7413     -2.9613     2.3712
   12     .          .*        .      31.42    29.2977      2.1223     2.6746
   13     .    *     .         .      17.74    27.6454     -9.9054     2.7932
   14     .          . *       .      35.99    31.1982      4.7918     2.1940
   15     .          .  *      .      39.07    32.2059      6.8641     1.5409
   16     .          .    *    .      42.91    30.1800     12.7300     2.3790
   17     .        * .         .      29.60    31.0481     -1.4481     1.8426
   18     .          .   *     .      40.87    32.4499      8.4201     1.4313
   19     .       *  .         .      27.13    32.0845     -4.9545     3.4069
   20     .        * .         .      34.00    36.2326     -2.2326     2.7588
   21     .          *         .      34.97    34.2759       .6941     2.7753
   22     .          .*        .      28.62    26.9459      1.6741     3.4442
   23     .          .  *      .      34.43    28.8539      5.5761     3.2488
```

FIG 4.5.2.5(e) Continued

```
              -3.0        0.0        3.0
Case #         0:........:.........:0        Y2        *PRED        *RESID    *SEPRED

      24       .        * .          .     23.17      28.5032       -5.3332    2.4888
      25       .         *.          .     25.06      27.7832       -2.7232    2.4954
Case #         0:........:.........:0        Y2        *PRED        *RESID    *SEPRED
              -3.0        0.0        3.0
```

Table 4.4.2.5(a) shows the ANOVA table with significance $F = 0.0357 \quad 0.05$, although the t - values for the variables X_1, X_3 and X_4 has p - values $p \quad 0.05$.
Table 4.4.5(b) shows the predicted values of Y_2 for each observed value of Y. Table 4.4.2.5(c) shows the t - values and we observe that only X_2 is the only variable contributes significantly to the relationship of the variable to Y_2.
Figs 4.4.2.5(c), 4.4.2.5(d), 4.4.2.5(e) shows the model adequacy using the Normal probability plot, standardized scatter plot and plot of standardized residuals respectively.

$$\begin{bmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \\ \hat{\beta}_2 \\ \hat{\beta}_3 \\ \hat{\beta}_4 \end{bmatrix} = \begin{bmatrix} 8.31311 \\ 0.03712 \\ 0.28270 \\ 1.07586 \\ 0.42089 \end{bmatrix}$$

The least square fit is:

$$\hat{Y}_2 = 8.31311 + 0.03712X_1 + 0.28270X_2 + 1.07586X_3 + 0.42089X_4$$

$$\hat{Y}_2 = 8.31311 + 0.03712 \times 41 + 0.28270 \times 37 + 1.07586 \times 2.9 + 0.42089 \times 7.68$$

$$\underline{\hat{Y}_2 = 26.65}$$

The predicted total rainfall for 1985/86 rainy season given X_1, X_2, X_3 and X_4 is 26.65 inches (676.84mm)
A $100(1-\alpha)$ per cent predicted interval for the total rainfall for 1985/86 rainy season given X_1, X_2, X_3 and X_4 is

$$\hat{Y}_2 \pm t_{\alpha/2,\, n-p}\sqrt{MS_E\left[1 + X_o'(X'X)^{-1}X_o\right]}$$

$$26.65 \pm 2.086\sqrt{35.74301\,[1 + 0.000419]}$$

$$\underline{26.65 \pm 12.47}$$

The predicted total rainfall for 1985/86 rainy season is between 14.18 inches (360.17mm) and 39.12 inches (993.65mm).

CONCLUSION

This project investigated two major areas:

(a) Analysis of raw data

(b) Predictions of future observations

CHAPTER II illustrates two methods of showing rainfall distribution in any given year. Section 2.2 shows the box plots where year to year comparison is shown. Section 2.3 illustrates the constructed histograms in any given rainy season showing finer details of the distributions of the total rainfall in any season.

CHAPTER III and IV illustrate various procedures of predicting future observations. Section 3.4 shows future season predictions using a Truncated Exponential distribution. The results obtained in this section were not informative due to large variations in the variables used. The results obtained in this section are of no use to Agriculturalists, Designers and other water users. Section 4.4 and 4.5 illustrate within-season predictions. In this section the first part of the rainy season has been observed and the problem is to predict the remaining aspects of the rainy season. In this project we observed the first part of the 1985/86 rainy season. That is, we had the following variables X_1, X_2, X_3, X_4. The problem was to predict Y_1 and Y_2 of 1985/86 rainy season. Using Simple/Multiple Linear regression, we observed taht Y_1 is not significantly related to any of the independant variables under consideration in this project. Y_2 was significantly related to some of the independant variables under consideration. The results obtained in section 4.5 shows the predicted total rainfall for 1985/86. The results obtained in predicting Y_2 from any of X_o's or the combination of these X_o's generally give the same predicted values of Y_2. All 100 $(1 - \alpha)$ percent predicted intervals established in section 4.5 for both simple/multiple linear regression analysis traps the ACTUAL total rainfall for 1985/86 rainy season.

REFERENCES

1. Bailey M: On Extreme Daily rainfall in Zambia, Meteorological Notes Series A No. 4 (1969).

2. Cochem J. and Franquin P: A Study of the Agro Climatology of the Semi - Arid Area South of the Sahara. FAO/UNESCO/WHO Tech. Bulletin (1967).

3. Das J. C.: Commencement of Sowing Rains in Zambia, Meteorological Notes Series A No. 9 (1979).

4. Das J. C. : Analysis of Lusaka (City Airport) Rainfall, Hydrometeorological Report No. 8 (1974).

5. Das J.C.: The Estimation of Daily Probable Maximum Precipitation in Zambia Using Hashfied Techniques, Meteorological Reports No. 4 (1974).

6. Dooren J. and Pruitt W.O.: Crop Water Requirements: FAO Irrigation and Drainage Paper No. 24, FAO Rome, Italy (1977).

7. Hutchinson P: The Climate of Zambia, Zambia Geographical Association (1978).

8. Ismal S. A.: Meteorological Research Bulletin No. 1,2 Vol 5 (1979).

9. Lawless J. F.: Prediction Intervals for the Two-Parameter Exponential Distribution Technometrics 19 (1977).

10 Montgomery C. C. and Peck E. A: Introduction to Linear Regression Analysis, (1984). 3rd Ed. John Wiley and Sons, Inc.

11 Simango V. A. and Das J. C.: Use of Annual Rainfall Deciles to Assessing Drought in Zambia, Meteorological Notes Series A 17 (1979).

12 Wolf F. L.: Elements of Probability and Statistics, (1979 McGraw Hill.

13 The Gamma Distribution Applied to Rainfall in Zambia, Meteorological Notes Series A No. 9 (1971).

14 Production Guide, Zambia Seed Company, Lusaka, (1985).

APPEDIX A

VARIABLES UNDER CONSIDERATION IN THE PROJECT ARE AS FOLLOWS:

YEAR	VARIABLES					
	X_1	X_2	X_3	X_4	Y_1	Y_2
1960/1	52	46	6.17	15.12	160	27.67
1961/2	46	38	8.97	19.11	173	40.30
1962/3	52	47	5.54	17.98	162	43.16
1963/4	51	34	3.38	11.58	143	24.78
1964/5	53	38	2.98	10.08	143	23.50
1965/6	58	46	3.50	07.85	162	37.01
1966/7	74	48	1.40	10.00	180	31.28
1967/8	85	40	2.01	07.58	165	24.72
1968/9	65	50	5.64	20.10	174	36.53
1969/70	28	21	4.20	15.39	134	22.75
1970/1	40	45	5.90	14.72	155	31.78
1971/2	46	57	3.13	08.53	186	31.42
1972/3	78	32	2.62	06.27	176	17.14
1973/4	78	36	3.30	13.51	186	35.99
1974/5	52	60	4.03	13.61	175	39.07
1975/6	58	36	2.15	12.73	195	42.91
1976/7	68	26	2.96	13.89	169	29.60
1977/8	54	45	4.33	13.44	204	40.87
1978/9	61	47	2.48	17.76	158	27.13
1979/80	65	41	6.63	16.97	197	34.00
1980/1	69	32	6.12	13.43	212	34.97
1981/2	91	32	2.20	05.87	143	28.62
1982/3	43	51	3.19	07.17	140	34.43
1983/4	79	24	1.50	11.60	169	23.17
1984/5	71	49	1.02	10.94	173	25.06

APPENDIX B

RAINFALL DISTRIBUTION AT CHING'OMBE MISSION

YEAR	SEASON BEGIN	SEASON END	TRF	TRF FOR Q_1 DAY	TRF FOR Q_2 DAY	TRF FOR Q_3 DAY
1960/1	52	160	27.67	6.92	13.84	20.76
				74	93	120
1961/2	46	173	40.30	10.08	20.15	30.23
				76	93	131
1962/3	52	162	43.16	10.79	21.58	32.37
				75	121	137
1963/4	51	143	24.78	6.20	12.39	18.59
				70	103	127
1964/5	53	143	23.50	5.88	11.76	17.64
				85	106	128
165/6	58	162	37.01	9.25	18.50	27.75
				109	130	147
1966/7	74	180	31.28	7.82	15.64	23.46
				86	106	142
1967/8	85	190	24.72	6.18	12.36	18.54
				105	121	140
1968/9	65	174	36.53	9.13	18.26	27.39
				83	110	139
1969/70	28	134	22.75	5.69	11.38	17.07
				70	86	112
1970/1	40	155	31.78	7.94	15.88	23.82
				72	95	111
1971/2	46	186	31.42	7.90	15.80	23.70
				89	122	151
1972/3	78	176	17.74	4.44	8.44	13.30
				102	105	135
1973/4	78	186	35.99	9.00	18.00	27.00
				99	135	150
1974/5	52	175	39.07	9.77	19.54	29.31
				75	113	148
1975/6	58	195	42.91	10.73	21.46	32.19
				83	123	157

YEAR	SEASON BEGIN	SEASON END	TRF	TRF FOR Q_1 DAY	TRF FOR Q_2 DAY	TRF FOR Q_3 DAY
1976/7	68	169	29.60	7.40	14.80	22.20
				85	110	147
1977/8	54	204	40.87	10.22	20.44	30.66
				90	109	160
1978/9	61	158	27.13	6.78	13.56	20.34
				74	85	132
1979/80	65	197	34.00	8.50	17.00	25.38
				88	118	152
1980/1	69	212	34.97	8.74	17.49	26.23
				108	136	157
1981/2	91	143	28.62	7.15	14.31	21.41
				102	121	135
1982/3	43	140	34.43	8.61	17.21	25.82
				103	119	136
1983/4	79	169	23.17	5.79	11.58	17.38
				87	124	142
1984/5	71	173	25.06	6.26	12.53	18.80
				91	116	127

RAINFALL SUMMARY

STATION: CHINGO'MBE MISSION

YEAR : 1960/1

DAY	JULY	AUG.	SEPT.	OCT.	NOV.	DEC.	JAN.	FEB.	MAR.	APRIL	MAY	JUNE
1							0.25	0.07				
2							0.12	0.14		0.44		
3					0.08	0.12						
4							0.94					
5								0.30	0.20			
6							0.47	0.06	0.86			
7					0.81	0.34	0.13	0.43	0.04		0.04	
8									0.39			
9							1.25		0.34			
10							1.25	0.12				
11						0.03	0.32	0.44				
12			0.02			0.38			2.55			
13						1.57	0.15					
14					0.10			0.19	0.83			
15					0.30	0.40	0.63	0.70				
16					0.10		0.09	0.31				
17						0.97	1.26	0.13		0.53		
18				0.05			0.03					
19				0.02				0.27	0.30			
20				0.07		1.90						0.03
21					1.88							
22					0.74	0.11			0.01			
23					0.20			0.57	0.01			
24							0.25	0.53				
25					1.67			0.42				
26							1.37	0.04	0.13			
27						0.79	0.05	0.06	0.04			
28						0.59	0.24	0.03				
29					0.15	0.35			0.97			
30						0.53	0.15					
31						0.27	0.04					
TOTAL	NIL	NIL	0.02	0.14	6.03	8.95	7.51	4.74	7.37	0.97	0.04	0.03
No. of Days of rain			1	3	10	13	18	18	14	2	1	1

All records with 2 decimal places implies inches
All records with 1 decimal place implies mitre