APPLICATION OF COMPUTER PROGRAMMING TO REGRESSION AND CORRELATION ANALYSIS

(A Case Study of National Cereal Research Institute, Baddegi)

By

Isyaku Ibrahim

PGD/MCS/96/020

A Project Submitted to the Post Graduate School, Department of Mathematics and Computer Science, Federal University of Technology, Minna, Niger State, Nigeria.
In partial fulfilment of the requirements for the award of Post Graduate Diploma in Computer Science.

March, 98

CERTIFICATION.

I Certify that this work was carried out by ISYAKU IBRAHIM (PGD/MCS/96/020) of the School of Post Graduate Studies, Federal University of Technology, Minna.

Dr. Yemi Aiyemisi
Project Supervisor

Date

Prof. K.R. Adeboye
Head of Department

Date

External Examiner

Date

DECLARATION

I Mr. Isyaku Ibrahim (PGD/MCS/96/020) hereby declare that this project has been written by me and that it is a record of my own research work. It has not been presented before in any form to any other organization for the purposes of Post Graduate Diploma or Higher Degree. All Sources of information as appropriate are acknowledged by means of references.

DEDICATION

With gratitude to ALLAH, this work is dedicated to my parents Alhaji and Mrs O. Opeyemi and all those who shared the pains of the period of running this programme of higher studies.

ACKNOWLEDGEMENT

I will seize this opportunity to express my profound gratitude to all those who in one way or the other contributed immensely towards my Academic success as well as the success of this project.

I am mostly indebted to God Almighty and my supervisor Dr.Yomi Aiyesimi whose cooperation and meaningful directives made this work a susccess. I also wish to express my gratitude to the Head of Depaartment (H.O.D.) of Mathematics and Computer Science Prof. K.R. Adeboye and to all my lecturers in the Department of Mathematics and Computer Science at the Federal University of Technology Minna. I also wish to express my gratitude to my parents Alhaji and Mrs Wahab Ogunleye and my friend Mr. Philemon, Student Centre, College of Education, Minna Niger State and also Mr. Gbadebo.

ABSTRACTS

This project studies Regression and Correlation Analyses with a view to developing a Computer Program to perform all necessary computation and the empirical assessment of the regression model.

The Computer program is developed with the use of BASIC programming language.

The data is collected on the Rice production, Rainfall and Temperature at the National Cereals Research Institute, Baddegi for the period of 10 years (1985-1996). Multiple Regression Model is used in order to predict the probable Quantity of rice production that may occur at the end of 1997, as affected by Rainfall and Temperature.

The test for the significance of the parameters of the model is done using the Analysis of Variance.

Correlation Coefficient is used to test the strength of Linear relationship that may exist between Rainfall and Quantity of Rice produced for the period under study. It is also used to determine the strength of the relationship between the quantity of rice produced and temperature recorded for the period.

The t-distribution is used to test for the multiple regression coefficient by estimating the coefficient of multiple determination R^2.

The Algorithm of the program is the Flowchart.

TABLE OF CONTENTS

CHAPTER THREE

SYSTEM ANALYSIS AND DESIGN

CHAPTER FOUR

PROGRAM DEVELOPMENT AND IMPLEMENTATION

CHAPTER FIVE

CHAPTER ONE

INTRODUCTION TO STATISTICAL DATA ANALYSIS.

It is obvious that not every one today is a statistician ,but every one, almost without choice ,must be a consumer of statistics. modern societies survive on data.Every business must know that the characteristics of its client ;intelligent consumers needs ,facts about competing products ;informed voters must have statistics on economic trend,voting record and population characteristics, sports fans, every where are masters of data consumption.

The scientific community relies on data even more,while the physical scientist is recording the readings of dials and gauges, the social scientist is counting people in the several categories,altering conditions under which people perform various events,and assessing the impact of the alteration.All these activities produce data which must be put into some kind of order before they yield the information hidden in these mass of numbers

DATA ANALYSIS

In broad sense, data analysis is the manipulation, summarization and display of data to make them more comprehensive to human minds, thus uncovering important departures from the structure of interest.

DATA COLLECTION

Often data relevant to a decision are not available either in internal records of the organisation or in external published sources.A survey can be carried out with respect to an entire population of interest or a small sample group taken from the population. In many cases sampling may be only logical way to find out something about a population. Data can also be obtained by direct observation, Interviews and questionnaires.

DATA SUMMARY

The most widely used method of summarising data is organising the data into a frequency table.

DATA DESCRIPTION

Methods employed to describe data are broadly divided into two parts.

(i) Measures of central tendency ;these include ,the mean, mode, median.

(ii) Measures of dispersion; these include,the range, variance, standard deviation and mean deviation.

MATHEMATICAL MODELS

Statistical procedures are derived from mathematical models of what is presumed to be reality, for example grading on a curve. This method is based on the belief that human talents are distributed normally, Many people cluster around average, A few are much above average and a similar sized group are very much below average.Mathematician then fit to this model of the world, a curve showing how proportion of the population will vary with talent level.Statistician in turn exploit this curve in describing groups of people and in making inferences about population.

Statistician use a variety of mathematical models of the world phenomena as foundation for developing their methods and procedures, one of such procedures are correlation and regression analysis which are discuses in the remaining part of this text.

1.2 COMPUTERS AND DATA ANALYSIS

The management challenge of huge social data files, especially census data, stimulated the development of several major innovations in computing. Automated data processing dates from 1890 US census, when the punched cards and tabulating machines invented by HERMAN HOLLEROITH were first used.

Holleroiths punched card system provided the foundation for electronic data processing (EDP),beginning with a unit record,electro-mechanical machines (EMM).

Statistical Analysis of data has been dramatically affected by computers. Infact the impact has been so great that it is difficult to imagine performing an adequate statistical analysis without the use of the computers.It continues to be one of the most active areas of continuing change and development.Statistical packages are expanding to address more of the tasks of data manipulation, storage, and presentation.

The look and feel of such programs is becoming more sophisticated with greater possibilities for user interaction, impressive graphic display and the beginning of machine intelligence. The future of this area promise to be as auspicious as its part.

(1.3) AIMS AND OBJECTIVES OF THE PROJECT.

The aim of this project topic includes the following.

i. To forecast the production of rice in the future using multiple regression analysis given by $Y= b_0 + b_i x_i + b^2 x^2$. Where Y represent the production rate X_i represent the temperature and X^2 represent the amount of rainfall.

ii. Testing for the significant of the parameter estimates in the multiple regression analysis using the analysis of variance. iii performing the t-test for the multiple regression coefficients, to test for the significant of b_o, b_1, b_2.

iv using the correlation coefficient to test for the relationship that exist between the parameters Y, X_1, X_2.

The objectives of this study will be

i. To know the effect of rainfall on rice yield.

ii. To know if temperature also has any effect on the production of Rice.

iii. Finally, to recommend necessary policy measures by which to maintain efficient rice production.

1.4 SCOPE AND LIMITATION.

The project is based on the application of computers programming to regression and correlation analysis (A case study of N.C.R.I).It is on the analysis of rice production,Rainfall and temperature from the year 1986-1996.

This will be limited to only the quantities of rice produced as it affected by rain and temperature.

CHAPTER TWO.

LIBRARY RESEARCH.

HISTORICAL BACKGROUND.

The National Cereals Research Institute (N.C.R.I.). was established in the year 1953 under the auspices of the then Department livestock and agriculture. This Institute is saddled with the responsibilities of undertaking basic research on crops (cereals) like maize, cowpea, bean seed, and rice with particular emphasis on RICE.

However, the institute was also changed with the responsibilities of seeing how well different varieties of cereal crops would growing Nigeria. When satisfied with the growth of these yield, the seed are normally distributed to other institution for multiplication and distribution of farmers for consumption.

The institute has five outstations on the country they are ,Mokwa in niger state, Ibadan in Oyo state, Amakama in Abia state, Birni-Kebbi in Kebbi state and in Kwara state .Its

Its head quarter was located at moor plantation but was moved to Baddegi, Niger state in 1984 during the Buhari\Idiagbon regime due to favourable soil and climate condition.

West Africa countries recently have been spending about Five years importing Rice, Demand has out stripped supply under existing production and market condition.Upper income consumer, largely in countries and Towns, have accounted for most increase in consumption by replacing other cheaper staple economic development. Total rice production amount to 1.5 million matric tons of paddy rice (called "rough rice "in the

U.S), on about 1.3 million hectares, Nigeria and Serria-leons are the largest producer accounting for one-half of the total, Senegal import two thirds of its rice supplies and accounts for one-half of all imports of its rice supplies and accounts for one-half of all imports in to the countries.

2.2 REGRESSION AND OTHER EXTENSIONS OF THE SIMPLE LINEAR REGRESSION MODEL.

There are various econometric methods that can be used to derive estimates of the parameters of economic relationship from statistical observations.

In this project we shall extend the simple linear regression model to relationships with two explanatory variables. Firstly we shall develop some practical rules for the derivation of the normal equations for models including any number of variables. Secondly we shall examine the extension of the two variable model to non-linear relationships.

MODEL WITH TWO EXPLANATORY VARIABLES.

THE NORMAL EQUATIONS.

We shall illustrate the three variable model with there given important to quantity of rice produced. The theory that the quantity of rice yield(Y) depends on the rainfall (X_1) and Temperature(X^2),

$$Y = F(X_1, X_2).$$

Given that the theory does not specify there mathematical form of the important function,we start our investigation by assuming that the relationship between Y, X_1, X^2 Is a Linear,

$$Y_i = b_0 + b_1 X1_i + b_2 X2_i + u_{i.}$$

Where i = 1, 2, 3............n.

This is an exact relationship whose meaning is that the variation in the role in which the two variables x_1 and x^2 play are fully explained by changes in amount of rainfall and temperature. If this form were true any observation on y, x_1, x^2 would determine a point which would lie on a plane. However, if we gather observation on these variables during a certain period of time and plot them on a diagram we will observed that not all of them lie on plane, some lies on it, but others lies above or below it. This dispersion is due to various factors, these are,

(1) Omitted from the function and to other types of error which are which are miss-interpreted.

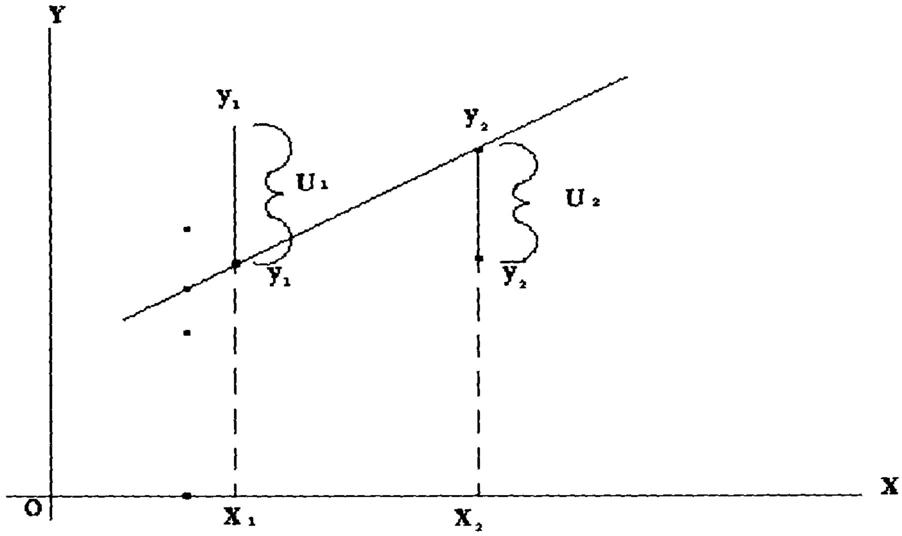
(2) Random behaviour of the human beings.

The dispersion of points around the line may be attributed to an erratic element which is inherent in human behaviour. This can be related to methods of planting and weeding.

(3) Imperfect specification of the mathematical form of the model. We may have linearised a possibly non linear relationship.

(4) Errors of aggregation. We often use aggregate data in which we add magnitudes referring to each months whose behaviour is dissimilar.

(5) Systematic Error.The deviations of points from the lines may be due to errors of measurement of the variables, which are inevitable due to the methods of collections and processing statistical information.



The true relationships which connects the variables involved is split in to two parts; a part represented by line and a part represented by then random term (u).

To complete the specification of our simple model we need some assumptions about the random variable (Y). These assumption include.

(1) The randomness of U. That is the variable U is a real random variable.

(2) Zero mean of μ. The random variable μ has zero mean value for each X_i.

(3) Normality of μ. The values of each (μ_i) are normally distributed.

μ_i---N(O,Qμ).

(4) Independent of μi and Xi. Every disturbance term Ui is independent of the explanatory variables.

E(UiXi) =E(UiXi) =0.

Having specified our model we next use sample observation on Y, X1, and X2 and obtain estimates of the true parameters b0,b1 and b2.

Yi=b0 + b1 X1 +b2X2.

Where b0,b1 and b2 are the estimates of the true parameters bo, b1 and b2 of the relationship. The estimates will be obtained by minimizing the sum of squared residuals.

$$\sum_{i=1}^{n} e^2 = \sum_{i}^{n} (y_i - y_i)^2 = \sum_{i}^{n} (y_i - bo - b_i x_{1i} - b_2 x_{2i})^2 \quad n \quad n$$

A necessary condition for this expression to assume a minimum value is that its partial derivatives with respect to b0, b1 and b2 be equal y zero.

$$\frac{\sigma\Sigma(Yi-b0-b1X1i-b2X2i)2}{\sigma bo} = 0$$

$$\frac{\sigma\Sigma(Yi-bo-b1X1i-b2X2i)2}{\sigma b1} = 0$$

$$\frac{\sigma\Sigma(Yi - b0\ b1X1i - b2X2i)2}{\sigma b2} = 0$$

Performing the partial differentiations we get the following system of the normal equations in the three unknown parameters b0, b1 and b2.

$\Sigma Yi = n\,b0 + b1\,\Sigma\,X1i + b2\Sigma X2i.$

$\Sigma X1iYi = b0\,\Sigma X1i + b1\Sigma X1i + b2\Sigma X1iX2i.$

$\Sigma X2iYi = b0\,\Sigma X2i + b1\Sigma X1iX2i + b2\Sigma X2I.$

From the above equations we determine the values for b0, b1 and b2. Which can be estimated as.

$$b0 = Y - b1X1 - b2X2.$$

$$b1 = \frac{\Sigma(X1iYi)\,(X2i) - (\Sigma X2iYi)(\Sigma X1iX2i).}{(\Sigma x1i\Sigma x2i) - (\Sigma X1iX2i).}$$

$$b2 = \frac{(\Sigma X2iYi)(\Sigma X1i) - (\Sigma X1iYi)(\Sigma X1iX2i).}{(\Sigma X1i)(\Sigma X2i) - (\Sigma X1iX2i)2.}$$

(2.3) THE SQUARED MULTIPLE CORRELATION COEFFICIENT

, r2y, X1,X2.

When the explanatory are more than one we talk of multiple correlation. The square of the correlation coefficient is called the coefficient of multiple determination or squared multiple correlation coefficient, it is denoted as r2 this shows the percentage of the total variation of rice yield (Y)explained by the regression plane that is by changes in the Rainfall(x1) and the temperature (X2).

r2y.X1,X2 = ΣY2 =Σ(Y-Y)2Σy2=Σ(Y-Y)2.

$$= 1 - \frac{\Sigma e2}{\Sigma Y2.}$$

$$= \frac{\Sigma y2 - \Sigma e2}{\Sigma y2.}$$

We established that ei = yi - yi and yi= b1 X1i +b2X2i

the squared residuals are Σe2i = Σei(yi - yi). By substituting in the formula of r2y.X1,x2.

We get r2y.X1,x2. $= \frac{\Sigma yi2 - b1\Sigma y1X1 - b2\,\Sigma y1X2i.}{\Sigma yi2.}$

$$= \frac{b1\Sigma YiX1i + b2\Sigma Y1X2i.}{\Sigma yi2.}$$

The value of r2 lies between 0 and 1, the higher r2 the greater the percentage of the variation of Y explained by the regression plane, that is the better the goodness of fit of the regression plane to the sample observation the closer r2 to zero the worse the fit. The above formula for r2 does not take into account the loss of degrees of freedom from the introduction of additional explanatory variables in the function.

(2.4) VARIANCE OF THE PARAMETER ESTIMATES b0, b1 and b2.

The estimates b0 ,b1,and b2 are unbiased estimates of the true parameters of the relationship between Y,X1 and X2.

1. The standard error test we print the standard errors

($Sb_i = \sqrt{Var(b_i)}$ } underneath the respective estimated and compare them with the numerical values of the estimates.

a. If $S(b_i) > ½b_i$. We accept the hypothesis, that is we accept that the estimate b_i is not statistically significant at the 5 percent level of significant.

b. If $S(b_i) < ½b_i$ we reject the null hypothesis, in other words we accept that our parameter estimate is statistically significant at 5 percent level of significance for a two tail test.

2. The student's test of the null hypothesis we compute the t - ratio for each b_i

$$t^* = \frac{b_i}{S()b_i)}$$

With n - k = n-3 =10 -3 =7 degree of freedom. The theoretical values of t at the chosen level of significance are the critical values that define the critical region in a two tail test, with n -k degree of freedom.

a. If t^* falls in the acceptance region: That is if $-t_{0.025}$ (with n -k degrees of freedom) we accept the null hypothesis, that is we accept that b_i is not significant (at 5 percent level of significance) and hence the corresponding regressor does not appear to contribute to the explanation of the variations in y.

b. If t^* falls in the critical region we reject the null hypothesis and we accept that alternative one: b_i is statistically significant clearly the greater the value of t^* the stranger is the evidence that b^i is significant.

2.5 TESTING THE OVERALL SIGNIFICANCE OF A REGRESSION.

This test has been explained in the proceeding section for the multiple regression. The test aims at finding out whether the explanatory variables (x_1,x_2 ....x_k) do actually have any significant influence on the dependent variable formally the y test of the overall significance of the regressions implies testing the null hypothesis.

$H_0 : b_1 = b_2 = b_3 = 0$

against the alternative hypothesis

H_i: not all b_i's are zero

If the null - hypothesis is true, that is if all the true parameters are zero, there is no linear relationship between y and the regressors.The test of the overall significance may be carried out with the table of analysis of variance. We compute the regression of y on x_1 and x_2 and we estimate

a. The total sum of square deviations of the y's Σy^2;

b. The sum of squared deviations explained by all the regressors (x_1 , x_2) Σy^2.

c. The sum of residual deviations Σe^2. From these terms we can evaluate the expression $\Sigma y^2 = \Sigma y^2 + \Sigma e^2$. We next find the degrees of freedom for each of the terms of the identity. The degrees of freedom for Σy^2 is k -1 where k = (k+1) is the total number of b's including the constant intercept. The degrees of freedom for Σe^2 is N -k. Where N is the sample size. The degrees of freedom of the total sum of squares is (k-1) +(N-k) =N-1 with this information we may compute the F^* ratio as

$$F^* = \frac{\Sigma y^2/(k-1)}{\Sigma e^2/(N-k)}$$

Which is compared with the theoretical F(at the chosen level of significance) with $V_1 = k - 1$ and $V_2 = N - k$ degrees of freedom if $F^* > F$ we reject the null - hypothesis ie, we accept that the regression is significant: not all b_i's are zero. If $F^* < F$ we accept the null - hypothesis, that is we accept that the overall regression is not significant.

The above information is summarized in the table of analysis of variance.

ANOVA TABLE

Source of variation	Degree of freedom	Sum of squares	Mean sum of squares	Fc
Replicate sum of squares(SS_{Rx_1})	b-1	SS_R	MS_R	MS_R/MS_E
Treatment sum of squares(SST_{x_2})	t-1	SS_T	MS_T	MS_T/MS_E
Residual sum of squares(SS_E)	(b-1)(t-1)	SS_E	MS_E	
TOTAL	bt-1	TSS		

CHAPTER THREE

SYSTEM ANALYSIS AND DESIGN

This chapter describes the design og the computer program

3.1 INPUT DESCRIPTION

Four data files namely **YRSFILE, TEMFILE,RAIFILE, and RIFILE** were created in order top provide a convenient means of storing data sets, since data files can easily be read and updated by a basic program. The YRSFILE data was created for the actual monthly gross production of Rice.

3.2 THE PRACTICAL OPERATION OF THE PROGRAM & OPERATION

The two lines (10 - 20) stated the title of the program. The next two lines (40 - 50) dimensioned the variables into their appropriate sizes. THese were followed by twenty three lines (50 - 250) for creating and storing the four input data files. YRSFILE for the yearly data, TEM file for tenperature recorded for the years, RAIFILE for the amonut of Rainfall and RISFILE for the total production for that year.

STEP 1

Line numbers 260 - 790 were mean't to compute the multiple regression line fofr the data. In achieving this, the original data was sorted out and their means were obtained and the values (constant b_0, b_1, b_2) were equally computed and regression equation is printed using the appropriate formulae.

STEP 2

3.3 TESTING FOR THE SIGNIFICANCE OF THE PARAMETER ESTIMATES USING ANVOA.

Line number 800 - 1300 were mean't to test if there is any significance different between the data. T o achieved these the data obtained were sorted using the binary sorting method. WE calculate the sum of squares and mean sum of squares using the appropriate formulae. Lastly, a decision was taken, at a 5% level of significance.

STEP 3

3.4 PERFORMING THE T- TEST FOR THE MULTIPLE REGRESSION COEFFICENTS.

Line number 1310 -1322 is sed to test for the significance of the parameters used b_0, b_1, & b_2. We estimate the coefficient of multiple determination R^2. We also compute the sum of squares and mean sum of squares, test for the parameter individually by comparing the computed result with the tabulated value at 5% level of significance, lastly a decision is taken.

STEP 4

Finally, line numbers 1323 -2030 were meant to test for the types of relationship that exist betweent the parameters using correlation coefficient. We first compute the r for Y_0, and X_1, and $Y_0$and X_2, X_1 and X_2 and there relationship were obtained after which decisions were taken as to which is correlated to one another.

The entire program is saved using the name IBRO BAS and can be accessed at the QBASIC environment. It is equally listed in appendix A.

CHAPTER FOUR

PROGRAM DEVELOPMENT AND IMPLEMENTATION

4.1 PROBLEM SPECIFICATION

The task of this design is to develop a system or program that could accept regression analysis data, stores it or otherwise, process the data and generate result on such things as the equation of the model: Correlation matrix and analyze the variance of the regression coefficient as well . The system should also be able to estimate the correlation co-efficient and test the data.

4.2 ALGORITHM

After an understanding of the input and output, a rigorous description of these which covers almost all cases were written down and analyzed. The following algorithms were set up for the several data, operations to be performed, after considering several alternatives.

Algorithm (solve for bi's)

(Estimates the coefficients of the regression line)

A → X X : X Y /

for k → N - 1 Do

find the smallest j > - k such that aj, k ≠ 0 if no such j exists then out put "no unique solution, stop.

Else

Exchange the contents of rows j and k of A for j ← k + 1 ...N Do

xm ← a_j, k/a_k,k

For p ← k +1...N +1 Do

a_{ji} ← a_{j1} p --- xm * a_{k},p

End for p,j,k

if a, $_{n}$ = 0 then out put "No unique solution"

stop

Else start backward substitution

b_n ← a_n n+ 1 / a_n, n

for i ← N -1 ....1 Do

b_i ← $\dfrac{(a_{j}, n+1 - \Sigma\, a_{i,\, ibj}}{ai,i}$

End for i

out put solution set ß = [b_i]

End

ALGORITHM ESTIMATE

(estimates the ANOVA components, Sb_i eb_i F not Se, and R)

(Takes XY array as input)

For i ← 1...N Do

For j ← 1..K Do

$x_{i,i}$ ← Σx_p, i * xp_{ij}

End for j, i

SST ← Σy_i^2

SSR ← Σ_{bi} Σx_j, i y_i

SSR ← SST - SSR.

Inverse matrix /$a_{i,j}$ / into / $c_{i,j}$

For i ← 1... K

$Sb_i \leftarrow \sqrt{\frac{c_{j,i} * SSE}{n-k-1}}$

$tb_i \leftarrow \frac{b_i}{sb_i}$

End for i

$Fc \leftarrow \frac{\frac{SSR}{k}}{SSE/N\_K1}$

$Se \leftarrow \sqrt{\frac{SSE}{N-K-1}}$

End

4.3 CHOICE OF PROGRAMMING LANGUAGE

Q- Basic it is defined as know as the beginner all purpose symbolic instruction code. It is mostly used by those who are writing a program for the first or those who are new into the program writing. It is choosing as the programming language for the following reasons.

1. It's has only view concepts to learn and digest.
2. It has a design that facilitates the writing of programs in a style that is accepted as a good programming practice.
3. It is very easy to implement and compile
4. It has a good file management facility

4.4 SYSTEM JUSTIFICATION

By hand, the computations necessary to perform a regression analysis are at best tedious, even with the assistance of a hand operated calculator. Such tedium can be alleviated by the use of a digital computer. The computer only save time and energy but also prevents the inevitable cascade of errors that occur whenever one error, such as mispunching the key of a calculator, is committed and provides greater level of accuracy. Although there are full featured statistical package that can handle regression analysis problem, apart from being expensive they imposes heavy hardware requirements. For example statistical package for social sentences (SPSS) requires a 16 - bit microcomputer 7 Million bytes of fixed dist for programs and data.More over, in most cases only users with sound statistical background can make use of the available features. This program is designed with a view to getting rid of these problems.

CHAPTER FIVE

OBSERVATIONS, RECOMMENDATIONS AND CONCLUSIONS

5.1 OBSERVATIONS

It was observed that in the course of this study, the successful use of regression analysis depends heavily must on the person doing the modelling the analyst must select relevant predictor variables and decide on the conclusion reach when resting the adequacy of the regression model.

More over, adding more and more variables to a model can lead to the problem of " over Vatting". The goal of a good regression analysis is to find a simple model with a reasonable number of predicator variable that make sense.

5.2 RECOMMENDATION

For this system to function well and indeed to perform regression analysis these are the requirements on the type and amount of data needed.

1. All data element must be in numerical form that is real integer.

2. To obtain estimates of the regression parameters, the number of data points N, must exceed the number of the predicator variables k.

3. If any of the predicator variables must be fire cast prior to using regression equation, then, the forecast of the variable must be easy to obtain. Modelling y as a function of variable that are themselves difficult to estimate will only decrease the accuracy of the y, forecast.

4. To perform any diagnostics procedure we need to have N greater than or $N \geq k$ exceed k by a reasonable amount. The larger $N - k$ is the more accurate the estimates will

be since all infernal estimates and hypothesis tests in regression analysis are based on S_e that is sample standard error.

In view of these it is strongly recommended that further study into the area of non linearity in variables and parameters and an extension of the program to accommodate data transformation be carried out.

5.3 CONCLUSIONS

When using regression analysis as a prediction tool, the analyst must bear in mind.

1. The conclusion and inferences made from a regression line apply only over the range of data contained in the sample used to develop the regression line.

2. The fact that a significance relationship exist between two variables does not imply that one variable cause the other.Also a cause and effect relationship between two variables is not necessary for regression analysis to be used for prediction, what matters is that the regression model accurately reflects the relationship between the two variables and that relationship remains stables.

APPENDIX

5.4 The programme.

5.5 The Sample Result.

```
10 REM THIS PROGRAM USES REGRESSION AND CORRELATION ANALYSES TO
ANALYSE A SET OF DATA
20 REM IT IS WRITTEN BY IBRAHIM ISYAKU PGD/MCS/020/96 AND SUPERVISED BY
Dr. A. Yomi.
30 DIM X1(100), Y(100), X2(100), S(100), YR(100)
40 DIM E(100), F(100), G(100), L(100), M(100), N(100)
50 KEY OFF: CLS : LOCATE 12, 35: INPUT " ENTER NUMBER OF YEARS", YN
60 CLS : LOCATE 23, 1: PRINT "READING THE DATA FOR YEARS"
70 OPEN "I", #1, "YRFILE.BAS"
80 FOR I = 1 TO YN
90 INPUT #1, YR(I)
100 NEXT I
110 CLOSE #1: OPEN "I", #1, "RIFILE.BAS"
120 CLS : LOCATE 23, 1: PRINT "READING THE YEARLY RICE PRODUCTION DATA
WAIT.."
130 FOR J = 1 TO YN
140 INPUT #1, Y(J)
150 NEXT J
160 CLOSE #1: OPEN "I", #1, "RAFILE.BAS"
170 CLS : LOCATE 23, 1: PRINT "READING THE YEARLY RAINFALL DATA WAIT..."
180 FOR K = 1 TO YN
190 INPUT #1, X1(K)
200 NEXT K
210 CLOSE #1: OPEN "I", #1, "TEFILE.BAS"
220 CLS : LOCATE 23, 1: PRINT "READING THE YEARLY TEMPERATURE DATA
WAIT..."
230 FOR I = 1 TO YN
240 INPUT #1, X2(I)
250 NEXT I
260 REM USING REGRESSION ANALYSIS TO ANALYSE THE DATA IN THE FILES
CREATED
270 REM FINDING THE MULTIPLE REGRESSION LINE FOR THE DATA WHICH IS
GIVEN BY y = b0 + b1x1 +b2x2
280 SUMX1 = 0: SUMX2 = 0: SUMY = 0
290 FOR I = 1 TO YN
300 SUMX1 = SUMX1 + X1(I): SUMX2 = SUMX2 + X2(I): SUMY = SUMY + Y(I)
310 NEXT I
320 REM LET THE MEANS OF Y , X1 AND X2 BE DENOTED BY A , B AND C
RESPECTIVELY
330 A = SUMY / YN: B = SUMX1 / YN: C = SUMX2 / YN
340 FOR I = 1 TO YN
350 P(I) = Y(I) - A: Q(I) = X1(I) - B: R(I) = X2(I) - C: E(I) = (P(I)) ^ 2: F(I) = (Q(I)) ^ 2
360 G(I) = (R(I)) ^ 2: L(I) = P(I) * Q(I): M(I) = P(I) * R(I): N(I) = Q(I) * R(I)
370 NEXT I
380 SUMP = 0: SUMQ = 0: SUMR = 0: SUME = 0: SUMF = 0: SUMG = 0: SUML = 0:
SUMM = 0: SUMN = 0
390 FOR I = 1 TO YN
400 SUMP = SUMP + P(I): SUMQ = SUMQ + Q(I): SUMR = SUMR + R(I): SUME = SUME
+ E(I): SUMF = SUMF + F(I)
410 SUMG = SUMG + G(I): SUML = SUML + L(I): SUMM = SUMM + M(I): SUMN = SUMN
+ N(I)
420 NEXT I
430 W = (SUML * SUMG) - (SUMM * SUMN): V = (SUMF * SUMG) - ((SUMN) ^ 2): S =
(SUMM * SUMF) - (SUML * SUMN)
440 b1 = W / V: b2 = S / V: b0 = A - (b1 * B) - (b2 * C)
450 ' LPRINT : LPRINT "---------------------------------------------------------------"
460 ' CLS : LOCATE 1, 3: LPRINT "WORKSHEET FOR THE REGRESSION EQUATION
y=b0+b1x1+b2x2"
470 '; LPRINT "------------------------------------------------------------"
```

```
480 ' LPRINT TAB(3); "YEARS"; TAB(12); "RICE Y"; TAB(21); "RAIN X1"; TAB(30); "TEMP
X2"; TAB(39); "yi=Yi-A"; TAB(47); "x1i=X1i-B"; TAB(57); "x2i=X2i-C"
490 ' YR = 1987
500 ' O = 1
510 ' FOR I = 1 TO YN
520 ' LPRINT TAB(2); YR(I); TAB(13); Y(I); TAB(23); X1(I); TAB(30); USING "##.##"; X2(I);
TAB(40); P(I); TAB(49); Q(I); TAB(59); R(I)
530 ' O = O + 1
540 ' NEXT I
542 ' LPRINT
544 ' LPRINT
546 ' LPRINT
550 ' LPRINT TAB(3); "(yi)^2"; TAB(16); "(x1i)^2"; TAB(27); "(x2i)^2"; TAB(39); "yix1i";
TAB(47); "yix2i"; TAB(57); "x1ix2i"
560 ' FOR I = 1 TO YN
570 ' LPRINT TAB(3); E(I); TAB(16); F(I); TAB(27); USING "##.##"; G(I); TAB(39); L(I);
TAB(47); M(I); TAB(57); N(I)
'580 ' NEXT I
580  LPRINT "*******************************************************************"
581  LPRINT "...........SAMPLE RESULT....................."
582  LPRINT "*******************************************************************"
584 LPRINT
590 LPRINT "THE MEAN OF Y = ", A
600 LPRINT "THE MEAN OF X1 = ", B
610 LPRINT "THE MEAN OF X2 = ", C
620 LPRINT "b1 = ", b1
630 LPRINT "b2 = ", b2
640 LPRINT "b0 =. ", b0
642 LPRINT
644 LPRINT
650 LPRINT "THE REGRESSION EQUATION y =", b0; b1; "x1"; "+"; b2; "x2"
660 REM FINDING THE ESTIMATED RICE YIELDS
670 FOR I = 1 TO YN
680 Y2(I) = b0 + (b1 * X1(I)) + (b2 * X2(I))
690 NEXT I
700 LPRINT : LPRINT "----------------------------------------"
710 CLS : LOCATE 1, 3: LPRINT " ESTIMATED RICE YIELD REPORT"
720 LPRINT "------------------------------------------"
730 LPRINT
740 LPRINT TAB(15); "YEARS"; TAB(28); "ESTIMATED RICE YIELD"
750 YR = 1987
760 D = 1
770 FOR I = 1 TO YN
780 LPRINT TAB(14); YR(I); TAB(31); Y2(I)
790 NEXT I
800 REM TESTING FOR THE SIGNIFICANCE OF THE PARAMETER ESTIMATES IN THE
810 REM MULTIPLE REGRESSION USING THE ANALYSIS OF VARIANCE
820 FOR I = 1 TO YN
830 U1(I) = (Y(I)) ^ 2
840 NEXT I
850 SUMU1 = 0
860 FOR I = 1 TO YN
870 SUMU1 = SUMU1 + U1(I)
880 NEXT I
890 SST = SUMU1 - ((SUMY) ^ 2) / YN
900 SUMY2 = 0
910 FOR I = 1 TO YN
920 U2(I) = (Y2(I)) ^ 2
930 SUMY2 = SUMY2 + U2(I)
940 NEXT I
```

```
950 SSR = SUMY2 - ((SUMY) ^ 2) / YN
960 SSE = SST - SSR
970 REM FINDING THE ASSOCIATED DEGREES OF FREEDOM
980 REM LET DF1,DF2 AND DF3 REPRESENT THE ASSOCIATED
990 REM DEGREES OF FREEDOM FOR SST,SSR AND SSE RESPECTIVELY
1000 K = 2
1010 DF1 = YN - 1: DF2 = K: DF3 = YN - K - 1
1020 MSR = SSR / DF2: MSE = SSE / DF3
1030 F1 = MSR / MSE
1031 LPRINT : LPRINT "------------------------------------------------------------------"
1032 CLS : LOCATE 1, 3: LPRINT "THE TESTS FOR SIGNIFICANCE OF THE
PARAMETER ESTIMATES USING ANOVA"
1033 LPRINT "--------------------------------------------------------------------"
1034 LPRINT "H0: Bi = 0 (THERE IS NO SIGNIFICANT RELATIONSHIP BETWEEN Y AND
X1,X2)"
1035 LPRINT "H1: NOT ALL Bi's ARE EQUAL TO ZERO"
1036 LPRINT "THE LEVEL OF SIGNIFICANCE IS 0.05"
1040 LPRINT "SST = ", SST
1050 LPRINT "SSR = ", SSR
1060 LPRINT "SSE = ", SSE
1070 LPRINT "DF1 = ", DF1
1080 LPRINT "DF2 = ", DF2
1090 LPRINT "DF3 = ", DF3
1100 LPRINT "MSR = ", MSR
1110 LPRINT "MSE = ", MSE
1120 LPRINT "F1 = ", F1
1121 F(CRITICAL) = 4.74
1122 LPRINT " F(CRITICAL) AT 2 AND 7 DEGREES OF FREEDOM = ", F(CRITICAL)
1130 LPRINT : LPRINT TAB(10); "-----------------------------------"
1140 CLS : LOCATE 1, 3: LPRINT TAB(9); "THE ANOVA TABLE FOR THE DATA"
1150 LPRINT TAB(9); "-----------------------------"
1160 LPRINT
1170 LPRINT
1180 LPRINT : LPRINT TAB(3); "----------------------------------------------------------------"
1190 LPRINT TAB(3); "SOURCE OF"; TAB(18); "SUM OF"; TAB(33); "DEGREE OF";
TAB(48); "MEAN"; TAB(63); "F*"
1200 LPRINT TAB(3); "VARIATION"; TAB(18); "SQUARES"; TAB(33); "FREEDOM";
TAB(48); "SQUARE"
1210 LPRINT TAB(3); "--------------------------------------------------------------"
1220 LPRINT TAB(3); "REGRESSION"; TAB(17); SSR; TAB(37); K; TAB(47); MSR;
TAB(62); F1
1230 LPRINT TAB(3); "ERROR"; TAB(17); SSE; TAB(37); YN - K - 1; TAB(47); MSE
1240 LPRINT TAB(3); "----------------------------------------------------------"
1250 LPRINT TAB(3); "TOTAL"; TAB(17); SST; TAB(37); YN - 1
1260 LPRINT TAB(3); "---------------------------------------------------------"
1270 LPRINT
1280 LPRINT
1290 IF ABS(F1) >= F(CRITICAL) THEN LPRINT "WE REJECT H0 AND CONCLUDE THAT
RICE YIELD IS SIGNIFICANTLY RELATED TO THE AVERAGE RAINFALL AND AVERAGE
TEMPERATURE"
1300 IF ABS(F1) < F(CRITICAL) THEN LPRINT "WE ACCEPT H0 AND CONCLUDE THAT
RICE YIELD IS NOT SIGNIFICANTLY RELATED TO THE AVERAGE RAINFALL AND
AVERAGE TEMPERATURE"
1310 REM PERFORMING THE t TEST FOR THE MULTIPLE REGRESSION
COEFFICIENTS
1320 REM ESTIMATING THE COEFFICIENT OF MULTIPLE DETERMINATION R^2
DENOTED AS DR
1330 DR = ((b1 * SUML) + (b2 * SUMM)) / SUME
1340 Vb1 = MSE * (SUMG / V): Sb1 = (Vb1) ^ .5
1350 Vb2 = MSE * (SUMF / V): Sb2 = (Vb2) ^ .5
```

```
1360 K2 = (B ^ 2) * SUMG: H2 = (C ^ 2) * SUMF: Z = 2 * B * C * SUMN
1370 Vb0 = MSE * ((1 / YN) + (K2 + H2 - Z) / V): Sb0 = (Vb0) ^ .5
1380 REM PERFORMING THE t TEST FOR b1
1390 t(CRITICAL) = 2.365
1400 t0 = b0 / Sb0
1410 LPRINT : LPRINT "-----------------------------------"
1420 LPRINT "TEST FOR THE SIGNIFICANCE OF b0 REPORT"
1430 LPRINT "-------------------------------------"
1440 LPRINT
1450 LPRINT
1460 LPRINT "H0 : B0 = 0 "
1470 LPRINT "H1 : B0 IS NOT EQUAL TO ZERO"
1480 LPRINT
1490 LPRINT "Vb0 = ", Vb0
1500 LPRINT "Sb0=", Sb0
1510 LPRINT "t0=", t0
1520 IF ABS(t0) >= t(CRITICAL) THEN LPRINT "WE REJECT H0 AND CONCLUDE THAT
bo IS STATISTICALLY SIGNIFICANT AT THE 5% LEVEL"
1530 IF ABS(t0) < t(CRITICAL) THEN LPRINT "WE ACCECPT H0 AND CONCLUDE THAT
b0 NOT IS STATISTICALLY SIGNIFICANT AT THE 5% LEVEL"
1540 t1 = b1 / Sb1
1550 LPRINT : LPRINT "-----------------------------------"
1560 LPRINT "TEST FOR THE SIGNIFICANCE OF b1 REPORT"
1570 LPRINT "-------------------------------------"
1580 LPRINT
1590 LPRINT
1600 LPRINT "H0 : B1 = 0"
1610 LPRINT "H1 : B1 IS NOT EQUAL TO ZERO"
1620 LPRINT
1630 LPRINT "Vb1 = ", Vb1
1640 LPRINT "Sb1 = ", Sb1
1650 LPRINT "t1 = ", t1
1660 IF ABS(t1) >= t(CRITICAL) THEN LPRINT "WE REJECT H0 AND CONCLUDE THAT
b1 IS STATISTICALLY SIGNIFICANT AT THE 5% LEVEL"
1670 IF ABS(t1) < t(CRITICAL) THEN LPRINT "WE ACCECPT H0 AND CONCLUDE THAT
b1 IS NOT STATISTICALLY SIGNIFICANT AT THE 5% LEVEL"
1680 t2 = b2 / Sb2
1690 LPRINT : LPRINT "-----------------------------------"
1700 LPRINT "TEST FOR THE SIGNIFICANCE OF b2 REPORT"
1710 LPRINT "-------------------------------------"
1720 LPRINT
1730 LPRINT
1740 LPRINT "H0 : B2 = 0"
1750 LPRINT "H1 : B2 IS NOT EQUAL TO ZERO"
1760 LPRINT "Vb2 = ", Vb2
1770 LPRINT "Sb2 = ", Sb2
1780 LPRINT " t2 = ", t2
1790 IF ABS(t2) >= t(CRITICAL) THEN LPRINT "WE REJECT H0 AND CONCLUDE THAT
b2 IS STATISTICALLY SIGNIFICANT AT THE 5% LEVEL"
1800 IF ABS(t2) < t(CRITICAL) THEN LPRINT "WE ACCECPT H0 AND CONCLUDE THAT
b2 IS NOT STATISTICALLY SIGNIFICANT AT THE 5% LEVEL"
1810 LPRINT
1820 LPRINT "THE COEFFICENT OF MULTIPLE DETERMINATION DR  = "; DR
1830 REM USING CORRELATION ANALYSIS TO ANALYSE THE DATA
1840 REM FINDING THE CORRELATON COEFFICIENTS BETWEEN Y , X1 AND X2
1850 r12 = SUML / (((SUMF) ^ .5) * ((SUME) ^ .5))
1860 r13 = SUMM / (((SUMG) ^ .5) * ((SUME) ^ .5))
1870 r23 = SUMN / (((SUMF) ^ .5) * ((SUMG) ^ .5))
1880 REM FINDING THE PARTIAL CORRELATION COEFFICIENTS BETWEEN Y , X1
AND X2
```

```
1890 r123 = (r12 - (r13 * r23)) / (((1 - (r13) ^ 2) * (1 - (r23) ^ 2)) ^ .5)
1900 r132 = (r13 - (r12 * r23)) / (((1 - (r12) ^ 2) * (1 - (r23) ^ 2)) ^ .5)
1910 r231 = (r23 - (r12 * r13)) / (((1 - (r12) ^ 2) * (1 - (r13) ^ 2)) ^ .5)
1920 LPRINT : LPRINT "--------------------------------"
1930 LPRINT "THE CORRELATION COEFFICIENTS REPORT"
1940 LPRINT "--------------------------------"
1950 LPRINT
1960 LPRINT
1970 LPRINT "THE CORRELATION COEFFICIENT BETWEEM Y AND X1 DENOTED r12 =
"; r12
1980 LPRINT "THE CORRELATION COEFFICIENT BETWEEM Y AND X2 DENOTED r13 =
"; r13
1990 LPRINT "THE CORRELATION COEFFICIENT BETWEEM X1 AND X2 DENOTED r23
= "; r23
2000 LPRINT "THE PARTIAL CORRELATION COEFFICIENT BETWEEN Y AND X1
DENOTED r123 = "; r123
2010 LPRINT "THE PARTIAL CORRELATION COEFFICIENT BETWEEN Y AND X2
DENOTED r132 = "; r132
2020 LPRINT "THE PARTIAL CORRELATION COEFFICIENT BETWEEN X1 AND X2
DENOTED r231 = "; r231
2030 REM USING THE FISHER'S Z-SCORE TO TEST HYPOTHESES FOR THE TRUE
VALUES OF
2040 REM PARTIAL CORRELATION COEFFICIENTS
2050 LPRINT
2060 LPRINT "THE LEVEL OF SIGNIFICANCE OF THE TESTS IS 0.05"
2070 Z(CRITICAL) = 1.96
2080 Zr123 = .5 * LOG((1 + r123) / (1 - r123)): Zr132 = .5 * LOG((1 + r132) / (1 - r132))
2090 Zr231 = .5 * LOG((1 + r231) / (1 - r231)): Kn = 1: Sz = 1 / ((YN - Kn - 3) ^ .5)
2100 Z123 = Zr123 / Sz: Z132 = Zr132 / Sz: Z231 = Zr231 / Sz
2110 LPRINT : LPRINT "--------------------------------------"
2120 LPRINT "TEST FOR THE SIGNIFICANCE OF r123 REPORT"
2130 LPRINT "--------------------------------------"
2140 LPRINT
2150 LPRINT
2160 LPRINT "H0 : P123 = 0"
2170 LPRINT "H1 : P123 IS NOT EQUAL TO ZERO"
2180 LPRINT
2190 LPRINT "Sz = ", Sz
2200 LPRINT "Zr123 = ", Zr123
2210 LPRINT "Z123 = ", Z123
2220 IF ABS(Z123) >= Z(CRITICAL) THEN LPRINT "WE REJECT H0 AND CONCLUDE
THAT RICE YIELD AND AVERAGE RAINFALL ARE SIGNIFICANTLY CORRELATED
WHEN THE TEMPERATURE IS HELD CONSTANT"
2230 IF ABS(Z123) < Z(CRITICAL) THEN LPRINT "WE ACCEPT H0 AND CONCLUDE
THAT RICE YIELD AND AVERAGE RAINFALL ARE NOT SIGNIFICANTLY CORRELATED
WHEN THE TEMPERATURE IS HELD CONSTANT"
2240 LPRINT : LPRINT "--------------------------------------"
2250 LPRINT "TEST FOR THE SIGNIFICANCE OF r132 REPORT"
2260 LPRINT "--------------------------------------"
2270 LPRINT
2280 LPRINT
2290 LPRINT "HO : P132 = 0"
2300 LPRINT "H1 : P132 IS NOT EQUAL TO ZERO"
2310 LPRINT
2320 LPRINT "Zr132 = ", Zr132
2330 LPRINT "Z132 = ", Z132
2340 IF ABS(Z132) >= Z(CRITICAL) THEN LPRINT "WE REJECT H0 AND CONCLUDE
THAT RICE YIELD AND AVERAGE TEMPERATURE ARE SIGNIFICANTLY CORRELATED
WHEN THE RAINFALL IS HELD CONSTANT"
```

```
2350 IF ABS(Z132) < Z(CRITICAL) THEN LPRINT "WE ACCEPT H0 AND CONCLUDE
THAT RICE YIELD AND AVERAGE TEMPERATURE ARE NOT SIGNIFICANTLY
CORRELATED WHEN THE RAINFALL IS HELD CONSTANT"
2360 LPRINT : LPRINT "----------------------------------------"
2370 LPRINT " TEST FOR THE SIGNIFICANCE OF r231 REPORT"
2380 LPRINT "----------------------------------------"
2390 LPRINT
2400 LPRINT
2410 LPRINT "H0 : P231 = 0"
2420 LPRINT "H1 : P231 IS NOT EQUAL TO ZERO"
2430 LPRINT
2440 LPRINT "Zr231 = ", Zr231
2450 LPRINT "Z231 = ", Z231
2460 IF ABS(Z231) >= Z(CRITICAL) THEN LPRINT "WE REJECT H0 AND CONCLUDE
THAT RAINFALL AND TEMPERATURE ARE SIGNIFICANTLY CORRELATED WHEN THE
RICE YIELD IS HELD CONSTANT"
2470 IF ABS(Z231) < Z(CRITICAL) THEN LPRINT "WE ACCEPT H0 AND CONCLUDE
THAT RAINFALL AND TEMPERATURE ARE NOT SIGNIFICANTLY CORRELATED WHEN
THE RICE YIELD IS HELD CONSTANT"
2480 CLOSE #1, #2, #3, #4
2490 END
```

3.5 FLOWCHARTS

The following [illegible]ures represent program flow charts for the data entry procedure [illegible] computational procedures. The procedures represented by [illegible] is indicated below.

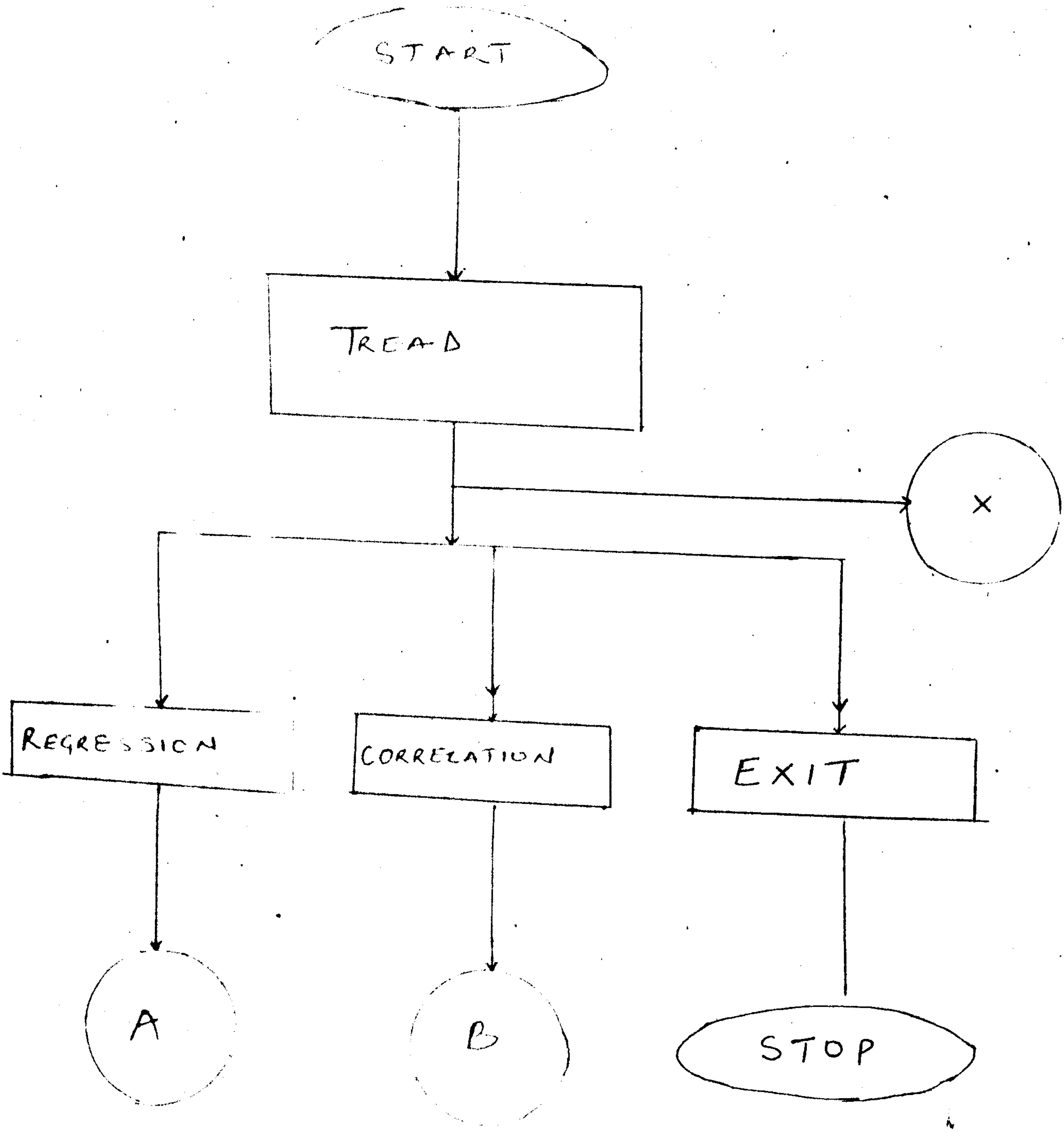


Fig: 3.5.1 Progr[illegible] structure.

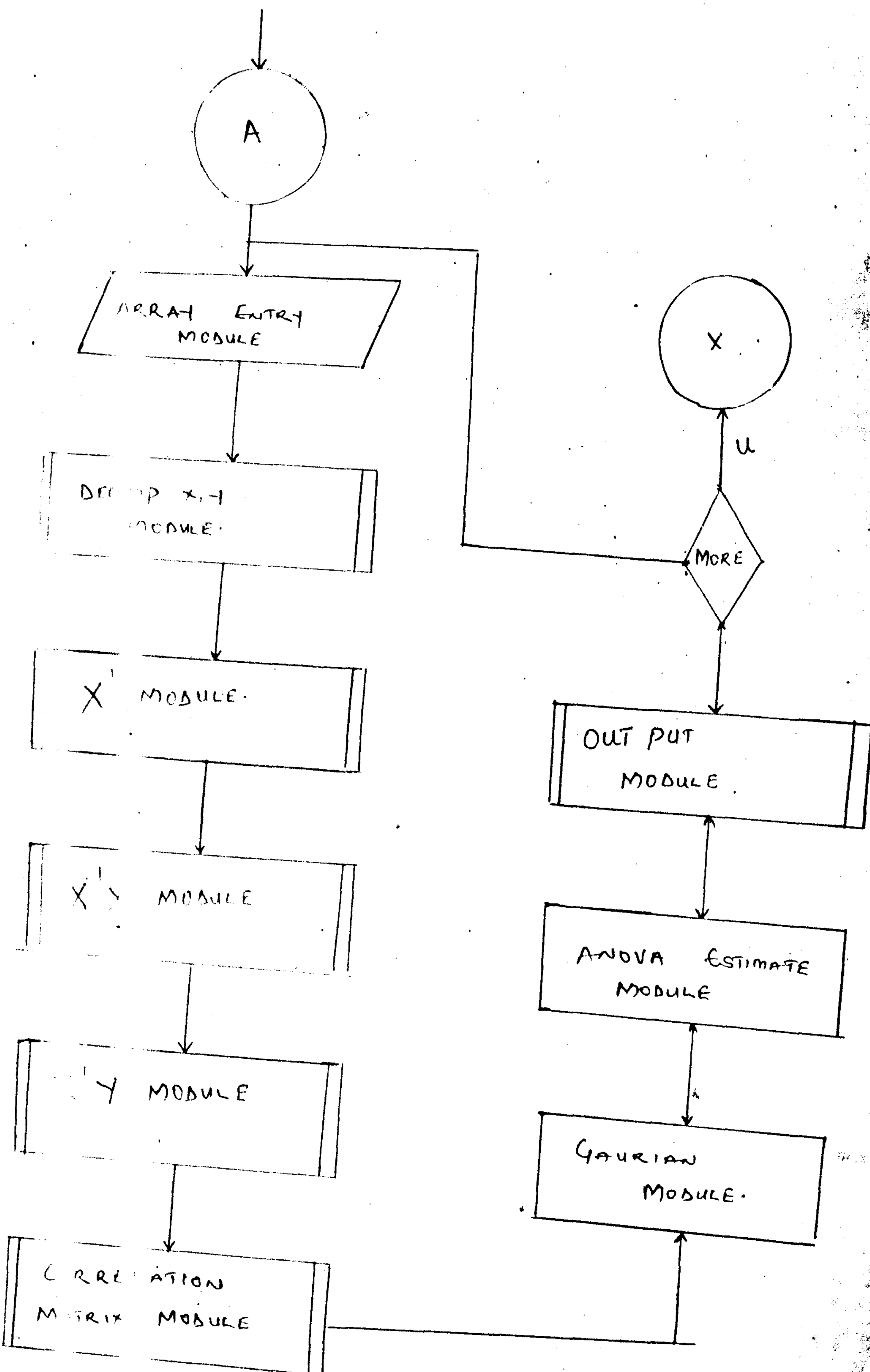


Fig. 3.5.1 continue.

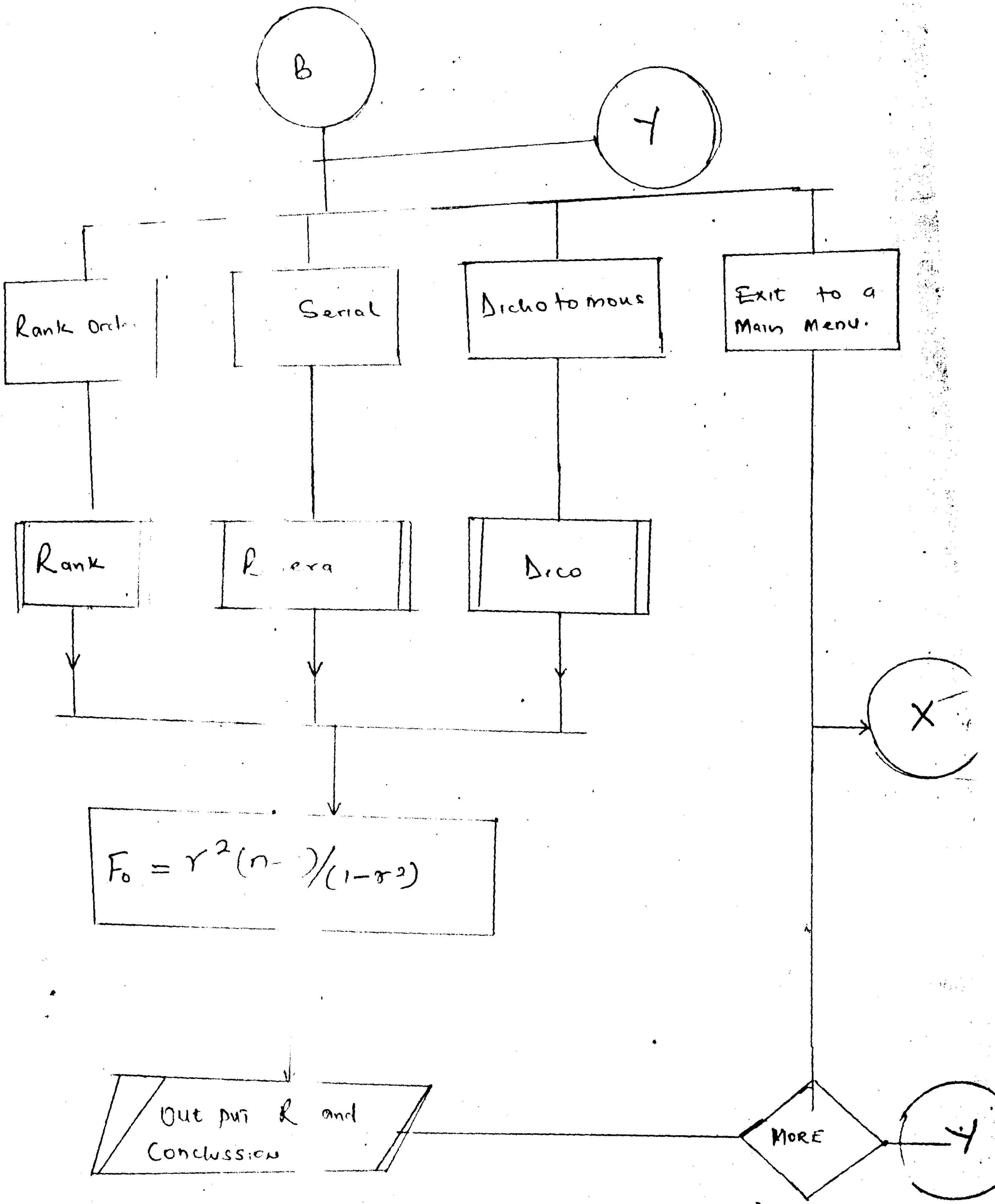


Fig. 3.5.1 continue.

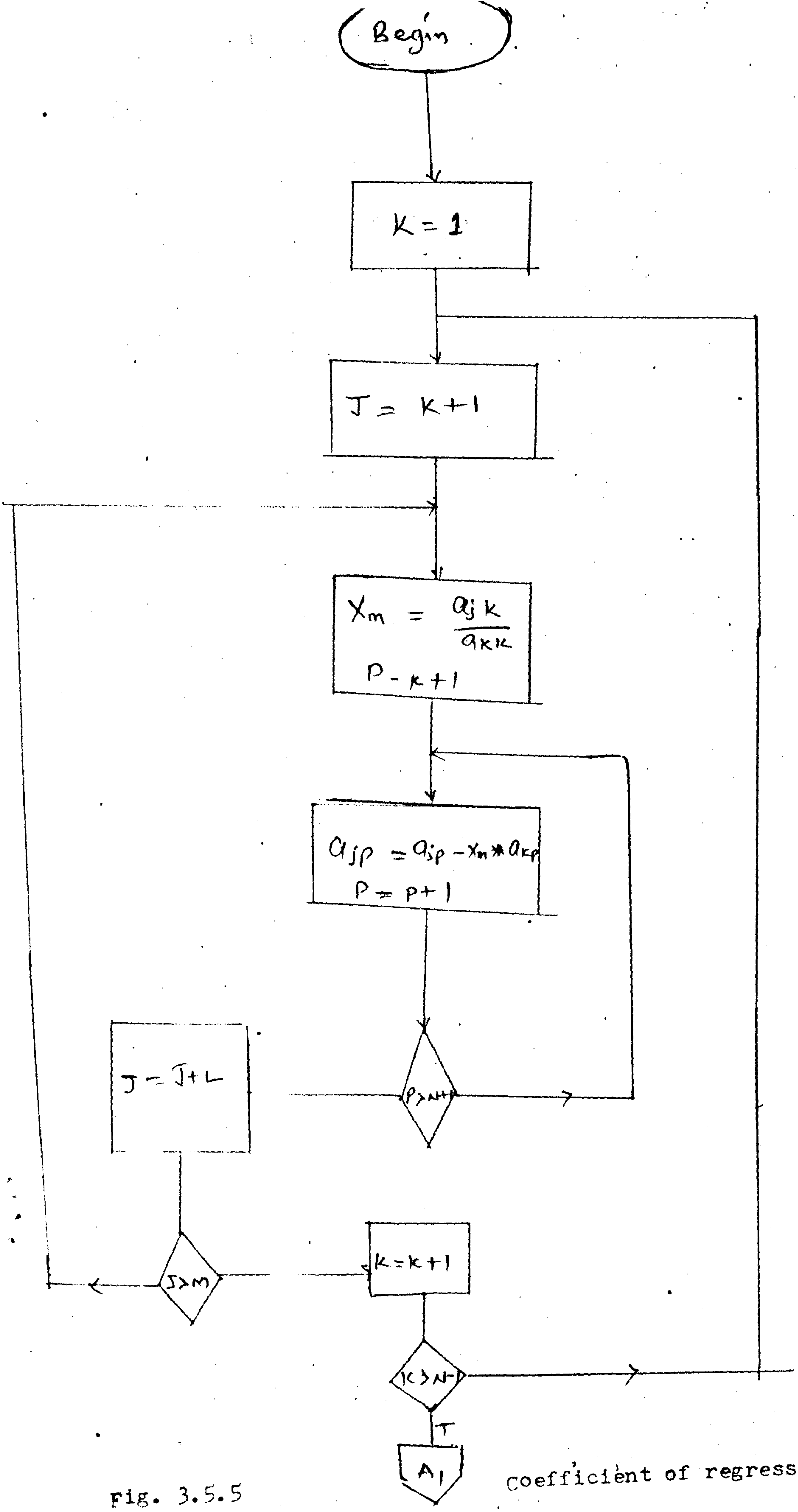


Fig. 3.5.5 Coefficient of regression estimate proce

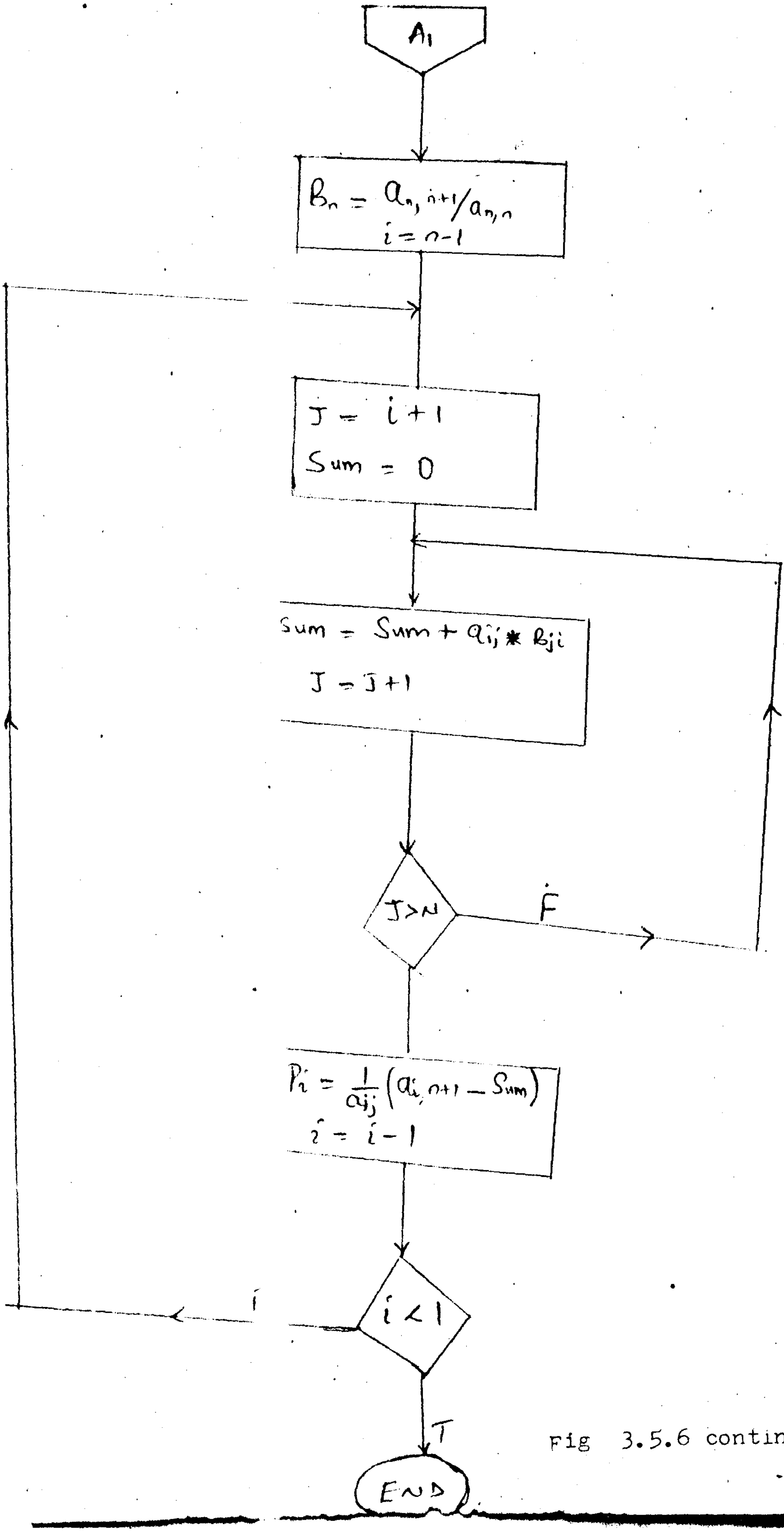


Fig 3.5.6 continue.

REFERENCES

1. Cochran W..G. (1976): Experimental Design John will and Sons (Publishiers) Ltd, New York.
2. Dennis. H. G. (May 1969): A computer Simulation Model for Selecting Advertising Schedules: Holden - Day (publisher) Ltd Sanfracesco, London.
3. Hadley G. (1968): Inconduction to Business statistics Holday - Day (Publisher) Ltd San Franasco London.
4. Henny J.C. (1970):Prediction of New product perfomance. After N.W. and Sons (Publisher) Ltd. New York.
5. ILERSIC A. R (1964) Econonmic and Social Statistics. HFL (Publishers) Ltd. 9 Bow Street Covent Garden London.
6. Montgnery D.C (1976) Design and Analysis of experiment. John Willey and Sons (Publishers) LTD New York.
7. Steel T. (1965): Principles of Statistics, Cron M.C Hill (Publisher) Convent Garden, London.
8. Went Z.B.W. (1972): Marketing Research Management and Methods. Harper and Row Publishers Inc New York.