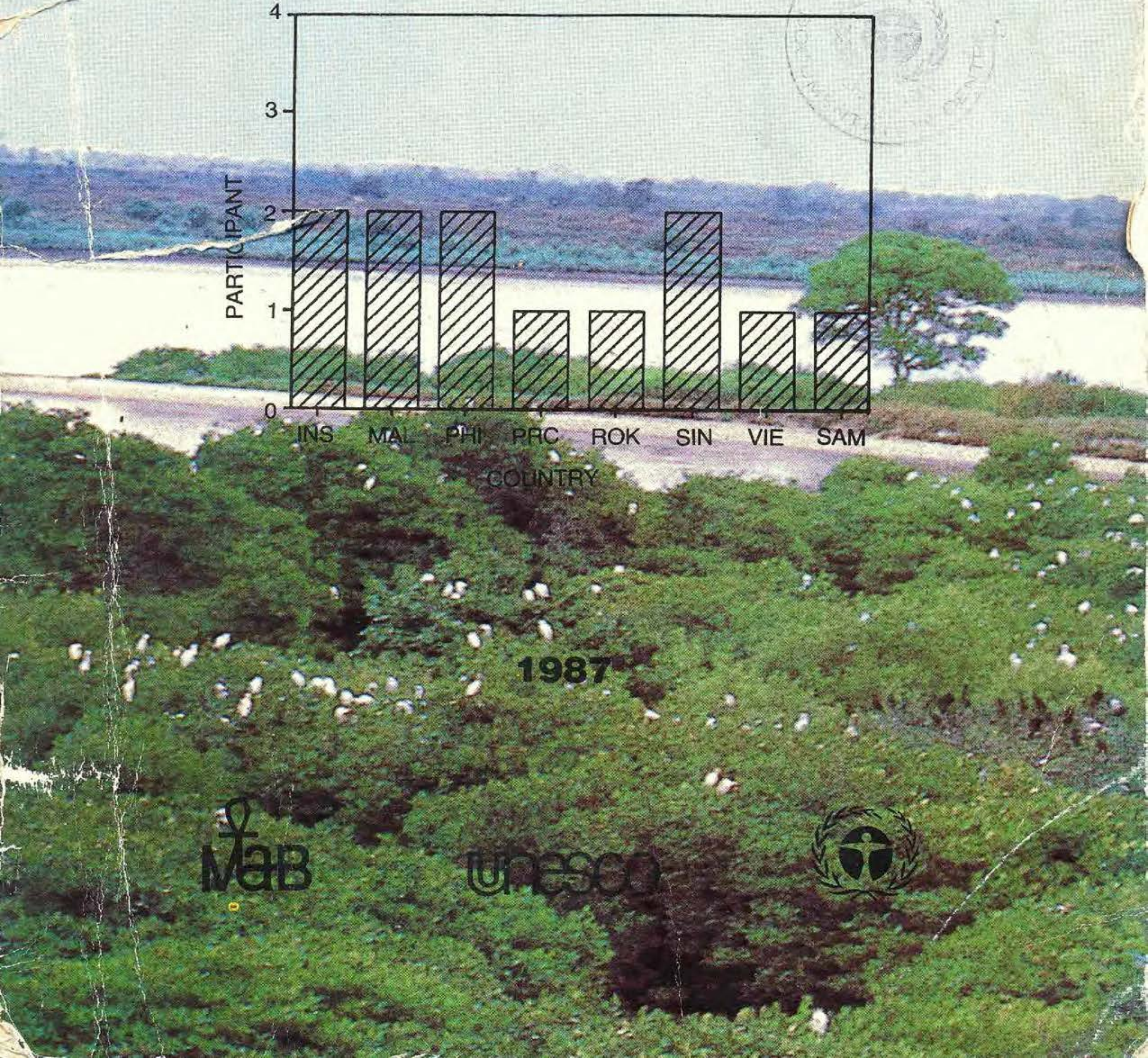
Computer Analysis for Environmental Biologists

Roger Green
Hong-Woo Khoo



1987

The cover photograph illustrator the mangrove ecosystem on the Pulau Dua Nature Reserve, West Jawa, Indonesia. Reproduced from on Indonesia MAB poster on "Cintaku Negeriku" series with permission.

ISBN. 92 - 3 - 102464 - 7

REPORT
ON
THE UNESCO-MAB/UNEP/NUS REGIONAL TRAINING COURSE
ON
COMPUTER - BASED QUANTITATIVE METHODS FOR ENVIRONMENTAL BIOLOGISTS

Singapore, 22 April – 11 May 1985

By

Roger Green
Department of Zoology
University of Western Ontario
London, Canada

Hong - Woo Khoo
Department of Zoology
National University of Singapore
Singapore

United Nations Educational, Scientific and Cultural Organization
Regional Office for Science and Technology
For Southeast Asia
Jalan M. H. Thamrin 14
Jakarta Pusat
Indonesia

1987

The authors of the papers published in this volume are solely responsible for the choice and presentation of the facts presented and for the opinions expressed therein. These are not necessarily those of UNESCO and do not commit the Organization in any way.

The designations employed and the presentation of material throughout the publication do not imply the expression of any opinion whatsoever on the part of UNESCO concerning the legal status of any country, territory, city or area, or of its authorities, or concerning the delimitation of its frontiers or boundaries.

CONTENTS

PREFACE

It is appropriate that a training course in computer-based quantitative methods for environmental biologists be held at this time and it is especially appropriate that it be held in Singapore. For developing countries, "development" must include the ability to use computers for many purposes, and one of the areas where computer skills are essential is certainly environmental biology. With the rapid and continuing increase in availability of microcomputers and softwares, there is no longer any barrier to the use of computer-based methods. Furthermore, it is advisable that training be given in use of state-of-the art systems rather than on the microcomputer systems of several years ago. Obsolescence is a problem in any area, but nowhere does it advance more rapidly than in the area of microcomputer technology and software development. We should not compound the problem by training for the use of out-of-date systems. In any case the newer systems and the software available to run on them is usually "simpler" for the user than the earlier systems were.

Singapore is an appropriate venue for such a course because as a country it has recognised the importance of the computer-based technological revolution. This is particularly true at the National University of Singapore where an excellent mainframe computer system and microcomputers linked in a network connected with the main system are available to students and staff. We hope that this training course will be just the first of a series of similar courses to be given in this region.

Roger Green

Hong-Woo Khoo

ACKNOWLEDGEMENT

We wish to express our appreciation to the Regional Office of Science and Technology for South East Asia, Unesco-Mab/Unep, for their support and to the Department of Zoology and the Computer Centre of the National University of Singapore for their facilities and assistance. In particular Miss Quek Lee Choo, course assistant, and Miss Tan Hiang Boon, should be acknowledged as well as Dr. J.R.E. Harger of Unesco-ROSTSEA for his support of the idea from the begining.

TRAINING COURSE ON COMPUTER-BASED QUANTITATIVE METHODS FOR ENVIRONMENTAL BIOLOGISTS

1. GENERAL INTRODUCTION

1.1 General objectives and approach

The emphasis is on biological modelling, especially the fitting of bivariate models and the use of them to test meaningful biological hypotheses.

Practical experience in application of the techniques is emphasized. Theory taught in lectures should be applied by doing the practical exercises. It should be emphasized that attendance at lectures without also taking time to do the practical exercises will be of limited value.

The biological modelling will be computer-based because it is the new micro and mainframe computer technology that allows biologists to effectively and efficiently model and analyze their data in ways that were impossible even a decade or two ago. However, this is not a "computer course", micro or mainframe, nor is it a course in computer programming. It is on modelling and it is aimed at biologists. A variety of computer hardware and software will be used to illustrate the diverse "tools" that are available to do biological modelling. But we shall attempt not to lose sight of the forest (biological modelling) as we become surrounded by trees (computer hardware and software).

The technology re. both hardware and software is advancing rapidly. You will always be out-of-date in the sense that someone, somewhere, has probably developed or is developing a better system for doing the job a biologist wants to do. Therefore present availability "back home" will not be the overriding criterion for the choice of hardware and software to be used in this course. In any country or region the system available now is not what will be available in a few years time, so that an attempt to train for use of present facilities would very soon become training for the past. Statistical packages that required large mainframe computers a few years ago now run on microcomputers. The "networking" of microcomputers so they can be used as terminals to access mainframe computers will more and more become common practice. This course will utilize a wide variety of hardware and software so that participants can gain experience with different systems, and then back in their countries they can help influence the direction of development of systems most useful to biologists. In summary, this course will train participants in use of hardware/software systems which will be available in the near future, even if they are not available in all countries now.

1.2. Structure of course

This report is based on a 3-week training course. Each day, Monday to Friday, was spent in 2 hr 15 min of lecture theory) in the morning and 3 hr of practical ("hands on") work in the afternoon. A library consisting of most of the references cited in the bibliography was available throughout the course. Presentations by participants were held on the one public holiday (May 1) and on the second Saturday.

2. INTRODUCTION TO HARDWARE AND SOFTWARE SYSTEMS

2.1 General remarks

As are noted in section 1.1, sophisticated languages and statistical packages are rapidly becoming available on microcomputers. One is no longer limited to the BASIC language, or to amateurishly written (sometimes incorrect) statistical programs. Most languages (FORTRAN, PASCAL, APL included) are now available for IBM PC-compatible systems, and some of them are available for APPLE-compatible systems. All of the major statistical packages used on mainframe computers (MINITAB, SAS, SPSS, BMD) are now available in "micro" versions, as well as a number of others especially designed for microcomputers.

However we are aware that many participants in the course, and readers of this report, are probably naive regarding languages other than BASIC, regarding use of statistical packages, and regarding use of mainframe computers. Therefore we have taken care to describe the general availability of the software we use in the course, in addition to commonly available hardware systems. Then we describe in great detail how to get started using the various hardware/software combinations.

The choices of languages and software are personal and subjective, and based on the senior author's experience. BASIC is ubiguitous on microcomputers and could not be ommitted. FORTRAN is the original profession scientific programming language, and many excellent programs are available. APL is not known as well as it should be, and is an excellent language for statistics and modelling. MINITAB is a good "friendly" beginning package for the most commonly used univariate statistical methods. SAS complements MINITAB by providing a wide variety of specialized methods (e.g., probit analysis) and multivariate methods (e.g., cluster analysis, principal components analysis, discriminant analysis). Other languages or packages might have been chosen, but these are appropriate ones.

2.2 Hardware & software Information

2.2.1. COMMENTS ON GENERALITY AND AVAILABILITY

Languages & Packages	OPERATING SYSTEM			
	Apple/DOS 3.3	Apple/CPM	IBM PC/MSPDOS	IBM mainfram/VM CMS
BASIC	This is the APPLE specific operating system. It is not on other micros. It uses a 6502 microprocessor. Some BASIC commands are APPLE-specific. 40 columns only.	This is a more general operating system, based on a Z80 microprocessor and available on a variety of micros. 80 columns.	The IBM-standard operating system BASIC is available in several versions. We will use MICROSOFT BASIC.	The VM/CMS operating system. own IBM BASIC.
FORTRAN		Available (but we will not run FORTRAN on the APPLE)	Available in several versions	Several versions of FORTRAN are on this system — we will use VS FORTRAN
APL			Available in several versions (we can run APL on the IBM PC)	IBM VS APL
MINITAB			Available for IBM PC with >256K RAM (but we will not use MINITAB on the PC)	Created orginally by Dept. of Statistics, Pennsylvania State University U.S.A.
SAS			Available for IBM PC with >256K RAM (but we will not run SAS on the PC)	Created originally by Dept. of Statistics North Carolina State University U.S.A.

2.2.2. REGRESSION AND SCATTERPLOT PROGRAM & PROCEDURES (+ NECESSARY UTILITIES)

Languages &	OPERATING SYSTEM			
Packages	Apple/DOS 3.3	Apple/CPM	IBM PC/MS DOS	IBM mainfram/VM CMS
BASIC	MAKE TEXT CREATE TEXT * REGRESSION PLOT *programs by Orloci and Kenkel	CREATE LINREG * *program CREATE by Green program LINREG by Somers	LINREG * *program by Somers	LINREG * * program by Somers
FORTRAN			REGR * PLOT *programs created or greatly modified by Green	REGR PLOT * * programs created or greatly modified by Green
APL				GLM * SCATTERPLOT *program GLM by Simillie, modified by Bailey program SCATTERPLOT by Anscombe, modified by Green.
MINITAB				READ and SET procedures. REGR procedure PLOT procedure
SAS				DATA step GLM procedure PLOT procedure

2.2.3. ANALYSIS OF COVARIANCE & MULTIVARIATE ANALYSES.

Languages & Packages	OPERATING SYSTEM			
	Apple/Dos 3.3	Apple/CPM	IBM PC/MS DOS	IBM mainframe/VM CMS
BASIC	Orloci & Kenkel programs: PCAR (PCA) ALC (avr link clustering) SSA (sum of sqrs clustering) WEIGHING/SCP (variable subset selection, as done by RSLCT FORTRAN)	ANOVA (by K. Somers, modified by R. Green)	ANOVA (by K. Somers, modified by R. Green	ANCOVA (by K. Somers, modified by R. Green)
FORTRAN				RSLCTIBM (by R. Green, from an algrithm by L. Orloci)
APL			APL operators can be used to do multivariate analyses	MATFORM R.Bailey COVAR PDET Ramsey & Musgrave ISOTROPY F. Anscombe GEIG R. Green and APL operators.
MINITAB				EIGEN, matrix algebra commands, and other commands.
SAS				PROC PRINCOMP PROC CLUSTER PROC CANDISC and other procedures.

2.3 Basic operation instructions

2.3.1 To start-up

2.3.1.1 IBM PC

Insert the DOS 2.0 diskette into drive A.
Switch on the computer switch at the right hand side.
Wait.
The following will appear:

```
A>wtdatim
Current date (DD-MM-YY):01-01-1980
Enter new date:   Press ENTER

Current time:00:01:00
Enter new time:   Press ENTER

A>REM The IBM Personal Computer DOS
A>REM Version 2.00 (C)Copyright IBM Corp 1981,1982,1983
A>
```

You can now proceed.

2.3.1.2 APPLE - DOS 3.3

Insert the DOS 3.3 diskette into drive A.
Switch on the computer switch at the back on the left side, and the monitor knob on top.
Adjust the switch attached to the right hand side of the monitor to the 40 column-number.
The following will appear:

```
DOS VERSION 3.3
APPLE II PLUS OR ROMCARD            SYSTEM MASTER

(LOADING INTEGER INTO LANGUAGE CARD)

]
```

Insert diskette containing required programmes into drive B.
Type in CATALOG, D2 to obtain a list of the files.

2.3.1.3 APPLE - CP/M

Adjust the switch on the right of the monitor to 80 column-number.

Insert the CP/M diskette into drive A.

Switch on the computer.

The following will appear:

```
Apple ][ CP/M
56K Ver. 2.20B
(C) 1980 Microsoft
A>
```

Insert diskette containing required programmes into drive B.

To obtain a list of the files in this diskette, type A>dir B:

2.3.1.4 MAINFRAME - IBM 3081 VM/CMS

The terminal and two of the IBM PCs are connected to the mainframe over at the Computer Centre.

Before you can use the system, you must have a user identification (userid) and password.

2.3.1.4.1 The TERMINAL

Switch on the terminal. After a short while, the logo:

```
NUS
VM/SP
```

should appear. Press ENTER key and logo disappears.

If, for example, your userid and password are DEMO1, log on to the system with the LOGON command, as follows:

```
Type  LOGON DEMO1
ENTER PASSWORD: DEMO1       Press ENTER
```

For security purposes, the password you enter is not displayed on the screen.

After logging-on, type the following:

```
CP DEF STOR 1500K          Press  ENTER
IPL CMS                    Press  ENTER
                           Press  ENTER A second time
```

Then you can carry on with XEDIT, MINITAB, SAS, BASIC, FORTRAN, APL, etc.

2.3.1.4.2 IBM PC used as Terminal

Insert the DOS 2.0 diskette into drive A and the IRMA Rev 1.10 diskette into drive B.
After starting-up, type:

```
                A>B:
and then        B>e78           Press ENTER
```

The NUS logo will appear. Log on as described above.

2.3.1.4.3 Some commands for IBM VM/CMS Operating System

LOGON acctname   The log-on procedure - the system will respond with a request for your password

DEF STOR 1500K   A request for temporary memory available to you to be increased to 1500K. This is necessary for running SAS, APL & certain other software. It is best to always do it. Following it you must return to CMS by entering 'I CMS' and depressing ENTER twice.

LIST   Lists all your files.

LIST fn ft   Lists a particular file, with name = fn and type = ft. For example, if you have a BASIC program in a file named 'REGR', then fn = REGR and ft = BASIC.

LIST fn   Lists all files with that fn. For example, if there is a file 'REGR BASIC' with a BASIC program in it and a file 'REGR DATA' with the data to be analysed in it, then both files would be listed if you entered 'LIST REGR'.

LIST * ft
Lists all files with that ft. For example, if you entered 'LIST * BASIC', then all files containing BASIC programs would be listed.

TYPE fn ft
Types the contents of that file on the terminal screen.

ERASE fn ft
Erases that file from your disk area. Use with care!

ALT + CLEAR
Clears the screen so the next screenful can be shown.

PRINT fn ft
Prints out the contents of that file (at the NUS Computer Centre).

LPRINT B08 fn ft
Prints out the contents of the file in the Computer Science Dept. lab room we will use (Comp Sci S15 02-11).

LPRINT C08 fn ft
Prints out the contents of the file at the Computer science Dept. printer that is operator-covered (one floor below 'our' lab room).

LOGOFF
The log off procedure (and switch off the terminal!).

2.3.2 To back-up files

2.3.2.1 IBM PC

After starting-up with DOS, type as follows:

```
A>diskcopy a: b:
```

The message will appear:

```
Insert source diskette in drive A:
Insert target diskette in drive B:
Strike any key when ready.
```

2.3.2.2 APPLE - DOS 3.3

Insert DOS 3.3 into drive A and empty diskette into drive B. To copy files, issue the command:

```
]RUN COPYA, D1
```

The following will appear:

```
APPLE DISKETTE DUPLICATION PROGRAMME

ORIGINAL SLOT: DEFAULT=6         Press RETURN
       DRIVE: DEFAULT=1          Press RETURN

DUPLICATE SLOT: DEFAULT=6        Press RETURN
        DRIVE: DEFAULT=2         Press RETURN

--- PRESS 'RETURN' KEY TO BEGIN COPY ---
```

2.3.3 Formatting a new disk

2.3.3.1 IBM PC

To format new diskettes:

```
A>format b:

Insert new diskette into drive B:
and strike any key when ready.
```

2.3.3.2 APPLE - DOS 3.3

To initialise new diskettes, insert new diskette into drive B.

```
]INIT HELLO, D2
```

2.3.4. To run a BASIC program

2.3.4.1 IBM PC

Start-up with DOS 2.0. Type in:

```
A>basic              Press ENTER
```

The screen will show:

```
The IBM Personal Computer Basic
Version D2.00 Copyright IBM Corp. 1981, 1982, 1983
61330 Bytes free

OK

_



1LIST  2RUN  3LOAD"  4SAVE"  5CONT  6LPT1  7TRON .....
```

You can now start running the program by typing in:

```
LOAD"  (name of the file)

RUN    (name of the file)
```

The file will be retrieved and the program will run.
To save time typing in the command LOAD and RUN, the function keys on the left of the keyboard can be used. With reference to the line printed at the bottom of the screen,

F3 is for the command LOAD

F2 is for the command RUN

2.3.4.2 <u>APPLE</u> - <u>DOS 3.3</u>

Type:

```
LOAD  (name of the file)

RUN   (name of the file)
```

2.3.4.3 <u>APPLE</u> - <u>CP/M</u>

Type: (note use of quotation marks)

```
LOAD " (name of the file)
RUN " (name of the file)
```

2.3.4.4 MAINFRAME - IBM 3081 VM/CMS

File with BASIC program in it must have filetype = BASIC
To go into BASIC mode:

```
basic
IBM BASIC/VM Version 1 Release 1.1.......
* run (name of the file)
```

To leave BASIC mode:

```
quit
```

2.3.5. To run a FORTRAN program on the MAINFRAME - IBM 3081 VM/CMS

File with FORTRAN program in it must have filetype = FORTRAN.
Data file to be used by program must have same filename, and filetype = DATA.
To run FORTRAN program that is in the file "fn FORTRAN":

```
fortvs fn
```

Lots of output follows, but the last three lines on the screen should be:

```
DMSLIO740I EXECUTION BEGINS.........
&EXIT
R;T=......
```

Your output is in file "fn OUTPUT". To see it, enter

```
type fn output
```

2.3.6 To run APL

2.3.6.1 IBM PC

Start up as before, using the Dos 2.0 diskette.
Insert the IBM APL diskette into drive B.
When A> appears, type B: to change drive.
Type:

```
B> APL
```

The following will appear

```
IBM Personal Computer APL
Version 1.00 (C) Copyright IBM Corp. 1983
Produced by IBM Madrid Scientific Center

CLEAR WS
```

Press the Control key Ctrl and the backspace key <--- (at the top row, right side of the keyboard), to invoke the APL character set. You may now proceed.

2.3.6.2 MAINFRAME - IBM 3081 VM/CMS

The terminal must have an APL character set.
Invoke the APL character set by depressing the ALT key and at the same time the <--- key that says "APL ON/OFF" on the front of it. The words APL should appear in the middle, under the horizontal line at the bottom of the screen.
To enter APL mode, type:

```
APL
```

To leave APL mode,

```
)OFF
```

2.3.7 MINITAB ON THE MAINFRAME - IBM 3081 VM/CMS

2.3.7.1. Some commands for using MINITAB

You have logged on and done 'DEF STOR 1500K'. To run MINITAB

(a) in interactive mode, enter 'MINITAB'. To get out of MINITAB enter 'STOP'. In this mode each entered command produces immediate response on the screen, but you can not get hard copy of your session -- not easily anyway.

(b) in batch mode, enter 'MINITAB fn' where the MINITAB commands are in a file named fn and with ft = MINITAB. The last command in that file must be 'STOP'. When the job has run, the output will be in a file with the same fn and ft = OUTPUT. You can use the 'PRINT' command to produce hard copy of those two files. (Or you can use 'LPRINT' - see section 2.3.1.4.3)

Please follow the following procedure in doing exercises. Do the complete exercise in interactive mode until you know you have it correct, exactly the way you want it. Write down the MINITAB cmmands that produce this perfect MINITAB run (including 'STOP' as the last command!), and enter them into a file (use XEDIT) with any fn you want (but it has to be ft=MINITAB). Then run it in batch mode, and then 'PRINT' out the 'fn MINITAB' and 'fn OUTPUT' files (see section 2.3.7.2 for more details).

Two useful things to remember about MINITAB commands are:

(a) MINITAB only pays attention to the first 4 letters of the command.

(b) MINITAB pays no attention to wards included in the command statement.

This means that you can enter 'REGRESS SIZE IN C1 ON 2 PREDICTOR VARIABLES TEMPERATURE IN C2 AND SALINITY IN C3, STORE STANDARDIZED RESIDUALS IN C4, PREDICTED VALUES IN C5'. Or you can enter 'REGR C1 2 C2 C3, C4, C5'. Both will work equally well. Why use the longer, wordier version? At least in the beginning it is better because:

(a) It will help you remember what you are doing and why, so the commands will make more sense and you will remember them more easily.

(b) When you go back to look at the hard copy, months or years later, it will be much easier to understand.

There will be MINITAB manuals around for you to refer to. Here are a very few commands to get you started:

REGR ---- you have just had this command described.

REGR C1 1 C2 would be the short version, for a simple linear regression if variable Y was stored in column CI and variable X in column C2. Neither residuals nor predicted Y-values would be stored.

READ C1-C3 would indicate that you will enter a 3-variable data set, one row at a time, and variables 1-3 will be stored in C1-C3, respectively.

SET C4 would indicate that you will enter a column of data, as a string of numbers, and it will be stored in C4.

PRINT C1-C4 would display the contents of C1-C4.

DESCRIBE C1-C4 would provide summary statistics for C1-C4 (number of elements, mean, standard deviation).

COPY C1 INTO C5 would copy the contents of C1 into C5, leaving C1 unchanged.

ERASE C1-C2 would erase the contents of C1 and C2 and leave them empty.

PLOT C1 VERSUS C2 would produce a scatterplot of the variable in C1 versus the variable in C2.

READ FROM 'fn' into C1-C3 would read data from a file named 'fn DATA' into columns C1-C3.

2.3.7.2 Sequence of steps for doing assignments in MINITAB

It is important that you follow this sequence:

1. Set up a blank sheet of paper as shown:
2. If you are going to read in data from a data file, then before entering the MINITAB interactive mode you must enter 'FI 8 DISK fn DATA(PERM)'.
3. Go into the MINITAB interactive mode. (Enter the command 'MINITAB'.)
4. Do the calculations step-by-step, using the MINITAB commands. In interactive mode you get an immediate response to each command. Look at each response to decide whether the correct command was entered. If it was, then write it down on your sheet under "Commands". Use the 'PRINT' command frequently (and write it down each time as well) to see what values are stored in columns, matrices, or constants, before they are used in commands or after they are produced by commands. (If you read in data from a data file, then you must use the 'DISKREAD' command.)

5. After you have done the entire assignment successfully in interactive mode, and have written down on your sheet all the commands needed to do an error-free MINITAB analysis run, then leave MINITAB interactive mode (enter 'STOP').
6. You are now back in CMS. Enter 'XEDIT fn MINITAB' (use whatever fn is appropriate), and you will go into the editor to create a file 'fn MINITAB'. Enter 'INPUT' and you will go into the INPUT mode within the editor. Enter the MINITAB commands that you wrote down when you did the assignment in interactive mode. (Remember to use READ rather than DISKREAD.)
7. When you have finished entering the MINITAB commands, leave INPUT mode by depressing 'ENTER' twice in succession without entering anything. Now you are out of the INPUT mode but still in the editor. Enter 'TOP' to see the file from its beginning. Carefully check what you have entered for errors, and correct any errors using the up, down, left, and right arrows. To get data from a data file ('fn DATA'), position the "active line" (brightly lit) on the "READ - -" statement line, and then enter 'GET fn DATA'. The data from 'fn DATA' will be inserted just below the READ -- statement line. Then enter 'FILE' to store the file and get out of the editor.
8. Now enter 'MINITAB fn'. The first response will say that your output is going into 'fn OUTPUT'. The second response (wait for it!) will be 'R;--'. Then continue.
9. Enter 'TYPE fn OUTPUT'. Examine your output on the terminal screen, making sure that it is the same results you obtained when you did the analysis in interactive mode. Notice that each command line or data entry line in the 'fn MINITAB' file produces dashes on a line in the 'fn OUTPUT' file. You may want to get rid of these lines. If so go into the editor again (by entering 'XEDIT fn OUTPUT'), delete the lines, and then enter 'FILE' to leave the editor.
10. Now you probably want a printed copy of both the MINITAB job command file ('fn MINITAB') and the output file ('fn OUTPUT'). Enter 'LPRINT B08 fn MINITAB' immediately

following that enter 'LPRINT B08 fn OUTPUT'

N.B.: You can not start an assignment involving MINITAB by attempting to create a 'fn MINITAB' file to use for a "batch mode" run!

2.3.7.3 MINITAB runs (interactive or batch) with file I/O

Before entering/running MINITAB the input and/or output data files, if there are any, must be identified. To identify an input data file, enter **FI 8 DISK fn DATA(PERM** and to identify an output data file, enter **FI 7 DISK fn DATA(PERM.**
The MINITAB command 'DISKREAD' (p. 34 of manual) must be used for input, and the command 'FPUNCH' (p.34 of manual) must be used for output.

It may be easier, when running MINITAB in batch mode, to follow the example given for running SAS in batch mode. That is, use XEDIT to incorporate the data file into the command file, and use 'READ' as if you were in interactive mode. But if you are running MINITAB in interactive mode, and are analysing a large data set that is in a file 'fn DATA', then you have little choice other than to follow the above instructions.

2.3.8 Some commands for using XEDIT (the editor on the IBM VM/CMS system)

XEDIT fn ft	Puts you into the editor, to edit the named file if it already exists, or to create it if it doesn't.
INPUT	Puts you into INPUT mode. Enter each line. When you depress the ENTER key, you are automatically given a new line to enter. To leave INPUT mode, depress the ENTER key twice in succession.

TOP

Moves the active line to the top line of the file.

BOTTOM

Moves the active line to the bottom line of the file.

UP 10

Moves the active line up 10 lines.

DOWN 15

Moves the active line down 15 lines.

PF8

Moves the active line down one screenfull.

PF7

Moves the active line up one screenfull.

up, down,
left and
right arrow
keys

Use these buttons to move the cursor around and make changes wherever you want. The changes are not stored until you depress the ENTER key the next time.

FILE

To leave the editor and store the file with all the changes you have made. Be careful! You are overwriting the file that existed when you went into the editor!

FILE fn ft

To leave the editor and store the file under the name fn and with the filetype ft. If you give a new fn, then you create a new file, and you do not overwrite the old file.

QQUIT

To leave the editor, abandoning all the changes you have made.

SAVE fn ft — To stay in the editor but save all the changes you have made so far.

GET fn ft — Inserts the named file into the file you are editing. It will be inserted just below the active line.

2.3.9 Some commands for running APL on the IBM VM/CMS system
(most of these commands also work on the IBM PC with APL)

In APL look at the red letters and symbols. Look at the black ones only where there is no red symbol.

2.3.9.1

)LOAD ws — Loads the named workspace from disk to your active area. This workspace will contain programs (called functions in APL) and variables (which can represent vectors or matrices of numbers in APL).

)FNS — This causes the functions in the active workspace to be listed.

)VARS — This causes the variables in the active workspace to be listed.

)WSID — This obtains the name of the active workspace (in case you forget it).

)SAVE — This saves the active workspace under the same name. Be careful! You are overwriting the workspace that you loaded from disk!

)SAVE ws — This saves the active workspace under the name ws. If you give a new name for ws, then you create a new workspace and do not overwrite the old one.

)ERASE n1 n2 n3 ---where 'n1 n2 n3 ---' are names of functions and/or variables. This erases the named functions and/or variables from the active workspace. (Before doing a 'SAVE' you should do 'FNS' and 'VARS' to see what garbage you have accumulated, and then do an 'ERASE' to get rid of it.)

vn

where vn is a variable name. The contents of the variable will be displayed. For example if the variable X contains the vector '1 2 3 4', and you enter 'X', then '1 2 3 4' will be displayed.

vn ← ----

This sets the variable named vn equal to whatever is to the right of the arrow. For example if you enter 'Y ← 4 6 7' and then enter 'Y', the response will be '2 4 6 7'. If X contains '1 2 3 4', and you now enter 'Z←X,Y' then Z will contain '1 2 3 4 2 4 6 7'.

vn3← vn1,vn2

Therefore, as just described, two vectors are put together. That is, they are "catenated".

vn3 vn1+vn2

This adds the two vectors, element by element. Obviously they must be the same length. For example if you entered 'Z← X+Y', then Z would contain '3 6 9 11'. Subtraction ('-') works the same way, and so does multiplication ('x') and division ('-'). <u>N.B.</u>: The minus sign for subtraction is at the upper right of the keyboard, whereas the "negative" sign to put in front of a negative number is at the upper left of the keyboard. They are different symbols in APL.

vn[3 4]

This will display the 3rd and the 4th elements of a vector whose name is vn. Obviously vn must be a vector, and it must have at least 4 elements. For example if you enter 'Y[3 4]'; the response will be '6 7' if Y contains '2 4 6 7'.

vn 4 2ρ ----

This creates a 4-by-2 data matrix from the vector of numbers represented by '----'. There must be 4x2 = 8 elements in the vector. For example if you enter 'D←4 2ρ 1 2 2 4 3 6 4 7', then D will contain the matrix '1 2'.

2 4
3 6
4 7

vn3← vn1,vn2

where vn1 and vn2 contain matrices rather then vectors. This catenates the matrices. For example if you enter 'D←X,Y' where X has been defined by 'X←4 1ρ 1 2 3 4', and Y by 'Y←4 1ρ 2 4 6 7', then D will contain the matrix '1 2'.

2 4
3 6
4 7

D[3;2]

This would result in the display of the '6' which is the element in the 3rd row and the 2nd column of D as defined above.

D[;2]

This would result in the display of the 2nd column of matrix D. If you entered 'D[;1 2]' or 'D[1 2 3 4;]', then you would get a display of all of matrix D. (That would be silly of course you could just enter 'D' and get the same thing. But if you enter 'D[3 4;]', then you will get a display of '3 6'.)
4 7

Y⌹XP

where 'Y←4 1⍴2 4 6 7', and 'XP←1,X' where 'X←4 1⍴1 2 3 4'. The result is a display of 0.5 and 1.7, the intercept and slope of the regression of Y on X.

)LIB

This displays the names of all your APL workspaces stored in your disk area.

)OFF HOLD

This causes you to leave APL mode, but stay logged on the system. (Be sure to SAVE your active workspace first if you have created something you want to keep!)

2.3.9.2 Examples of APL

```
ADDITION                                 MULTIPLY
       8 7 + 7 3                                  2 6 x 1 4

 15  10                                   2 24


SUBTRACT                                 DIVIDE
       4 6 - 2 3                                  2 6  - 1 4

 2 3                                      2 1.5


RECIPROCAL                               SHAPE
      ÷ 5 2                                       2 2⍴  1 2 3 4

 0.2  0.5                                 1 2
                                          3 4


ABSOLUTE VALUE                           TRANSPOSE
         3 ¯6 ¯5                   If A is     1 2  then     A
                                               3 4
3  6  5
                                          1 3
                                          2 4


NATURAL LOGARITHM                        CATENATE
          1 10                                    1 3 , 4 5

0  2.303                                  1 3 4 5
```

EXPONENT

```
      * 0 2.303

1  10
```

FACTORIAL

```
    !1 2 3 4

2 6 24
```

MATRIX INVERSE

```
if B is  2 1  then  ⌹B
         1 3

  0.6  ¯0.2
 ¯0.2   0.4
```

INDEXING

```
          [A 2;]

  3 4

           B[2;2]

  3
```

3. INTRODUCTION TO LINEAR REGRESSION

3.1. General remarks

Linear regression analysis is treated extensively in standard textbooks on introductory statistics and biometrics, some of which are cited in the bibliography. This section is not intended to substitute for a knowledge of regression analysis that should have been obtained from an introductory statistics course using such a textbook. Indeed, such a background was explicitly stated as a prerequisite for this course. This section is intended to be a review, and also to introduce participants to some new concepts and techniques. New concepts include matrix notation, matrix algebra, and derivation of the least squares regression formulae from first principles. New techniques include the use for regression analysis, of the software (BASIC, FORTRAN and APL languages; MINITAB and SAS statistical packages) and hardware (APPLE DOS 3.3, APPLE CP/M, IBM PC, and IBM 3081 mainframe) which will be used throughout the course.

3.2 A linear regression analysis on a small data set

X :	-2	0	+2	and n=3
Y :	+3	+1	-4	

(1) First we need the X and Y deviations, but in these data $\bar{X} = 0$ and $\bar{Y} = 0$ so the data are X and Y deviations.

Therefore

$$\Sigma Y^2 = \Sigma y^2 = 3^2 + 1^2 + (-4)^2 = 26$$
$$\Sigma X^2 = \Sigma x^2 = (-2)^2 + 0^2 + 2^2 = 8$$
$$\Sigma XY = \Sigma xy = (-2)(3)+(0)(1)+(2)(-4) = -14$$

(2) slope $b = \Sigma xy/\Sigma x = -14/8 = -1.75$

$\bar{y} = 0$

$\bar{x} = 0$

Therefore $\bar{Y} = a + b\bar{X}$

$0 = a + (-1.75)(0)$ and $a = 0$

The equation is $y = -1.75X$

(3) ANOVA of regression table:

Source	df	SS	MS	F
Regression	1	$(\Sigma xy)\ /\Sigma x^2$ $=(-14)^2\ /8$	24.5	24.5/1.5 = 16.3
Error	n-2 = 3-2 = 1	26-24.5=1.5	1.5	
Total	n-1 =3-1 =2	$\Sigma y^2 = 26$		

$$r^2 = \frac{\text{regr.SS}}{\text{tot. SS}} = \frac{(\Sigma xy)^2\ /\Sigma x^2}{\Sigma y^2} = \frac{(\Sigma xy)^2}{\Sigma x^2\ \Sigma y^2} = \frac{24.5}{26} = 0.942$$

$$r = \sqrt{r^2} = \sqrt{0.942} = 0.971$$

3.3 Examples of Matrix algebra.

Addition: $\begin{bmatrix} 1 & 5 \\ 5 & 4 \end{bmatrix} + \begin{bmatrix} 3 & 2 \\ 2 & 4 \end{bmatrix} = \begin{bmatrix} 4 & 7 \\ 7 & 8 \end{bmatrix}$

Subtraction : $\begin{bmatrix} 1 & 5 \\ 5 & 4 \end{bmatrix} - \begin{bmatrix} 3 & 2 \\ 2 & 4 \end{bmatrix} = \begin{bmatrix} -2 & 3 \\ 3 & 0 \end{bmatrix}$

Transpose : $\begin{bmatrix} 6 & 2 \\ 1 & 3 \end{bmatrix}' = \begin{bmatrix} 6 & 1 \\ 2 & 3 \end{bmatrix}$ (rows→columns)

Multiply : $\begin{bmatrix} 2 & 3 & 1 \\ 1 & 5 & 2 \end{bmatrix} \begin{bmatrix} 1 & 8 \\ 4 & 1 \\ 3 & 6 \end{bmatrix} = \begin{bmatrix} 17 & 25 \\ 27 & 25 \end{bmatrix}$ (rows by columns)

Inverse : $\begin{bmatrix} 2 & 1 \\ 1 & 3 \end{bmatrix}^{-1} = \begin{bmatrix} 0.6 & -0.2 \\ -0.2 & 0.4 \end{bmatrix}$

Check: $\begin{bmatrix} 2 & 1 \\ 1 & 3 \end{bmatrix} \begin{bmatrix} .6 & -.2 \\ -.2 & .4 \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$

"Divide": $\begin{bmatrix} 2 & 1 \\ 1 & 3 \end{bmatrix}^{-1} \begin{bmatrix} 5 & 2 \\ 2 & 6 \end{bmatrix} = \begin{bmatrix} 2.6 & 0 \\ -0.2 & 2.0 \end{bmatrix}$

(Note that the result is not a symmetric matrix, though the originals were)

Determinant: $\begin{vmatrix} 2 & 1 \\ 1 & 3 \end{vmatrix} = (2)\ (3) - (1)(1) = 5$

Roots and vectors: $\begin{bmatrix} 2 & 6 \\ .5 & 0 \end{bmatrix} = M$ Find roots and vectors of matrix M.

Solve the equation $|\lambda I - M|X = 0$,

which is
$$\left| \lambda \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} - \begin{bmatrix} 2 & 6 \\ .5 & 0 \end{bmatrix} \right| \begin{bmatrix} X_1 \\ X_2 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix} \qquad (1)$$

or
$$\left| \begin{bmatrix} \lambda & 0 \\ 0 & \lambda \end{bmatrix} - \begin{bmatrix} 2 & 6 \\ .5 & 0 \end{bmatrix} \right| \begin{bmatrix} X_1 \\ X_2 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix} \qquad (2)$$

or
$$\left| \begin{bmatrix} \lambda-2 & -6 \\ -.5 & \lambda \end{bmatrix} \right| \begin{bmatrix} X_1 \\ X_2 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix} \qquad (3)$$

Divide both sides by $\begin{bmatrix} X_1 \\ X_2 \end{bmatrix}$:
$$\begin{bmatrix} \lambda-2 & -6 \\ -.5 & \lambda \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix} = \begin{aligned} (\lambda-2)\lambda - (-6)(-.5) &= 0 \\ \lambda^2 - 2\lambda - 3 &= 0 \\ (\lambda-3)(\lambda+1) &= 0 \end{aligned}$$

So the roots are : $\lambda_1 = 3$ and $\lambda_2 = -1$

Note that $\begin{bmatrix} 3 & 0 \\ 0 & -1 \end{bmatrix} = \begin{bmatrix} \lambda_1 & \\ & \lambda_2 \end{bmatrix}$ and $M = \begin{bmatrix} 2 & 6 \\ .5 & 0 \end{bmatrix}$ are equivalent

For one thing, note that the sum of the diagonal is the same.

vector associated with each root can be found by substituting the root into the equation $\left[\lambda I - M\right]\left[X\right] = \left[0\right]$ (see equation (3) above). The vector for λ_1 is $\begin{bmatrix} X_1 \\ X_2 \end{bmatrix} = \begin{bmatrix} 6 \\ 1 \end{bmatrix}$ or any vector proportional to it.

To do matrix algebra using MINITAB (where M1 and M2 are matrices):

Addition :	'ADD M1 TO M2, PUT IN M3'
Subtraction :	'SUBTRACT M2 FROM M1, PUT IN M3'
Transpose :	'TRANSPOSE M1, PUT IN M2'
Multiply :	'MULTIPLY M1 BY M2, PUT IN M3'
Inverse :	'INVERT M1, PUT IN M2'
"Divide" :	'INVERT M1, PUT IN M2' and 'MULT M1 BY M2, PUT IN M3'
Determinant :	'EIGEN M1, C1, M2' & 'LET C2 = LOG E(C1)' and 'LET K1 = EXPO(SUM(C2))'
Roots & vectors:	'EIGEN M1, PUT ROOTS IN C1, VECTORS IN M2'

(N.B.): This will only work for a symmetric matrix. There is another way of doing it for a nonsymmetric matrix.

To do matrix algebra using APL (where vn1 and vn2 are matrices):

Addition :	'vn3← vn1 + vn2'
Substraction :	'vn3← vn1 - vn2'
Transpose :	'vn2←⍉vn1'
Multiply :	'vn3← vn1 +.x vn2'
Inverse :	'vn2←⌹ vn1'
"Divide" :	'vn3← vn2⌹ vn1'
Determinant :	Use function 'PDET vn'
Roots & vectors:	Use function 'GEIG vn'. Works for a symmetric or nonsymmetric matrix so long as the roots are fairly distinct (that is, none of the roots are approximately equal to each other).

3.4 Derivation of least squares regression formula

In matrix notation, $Y = XB + e$, where the e_i are independent estimates of , which (for t & F-tests) are normally distributed with 0 mean.

which is

$$\begin{bmatrix} Y_{i=1} \\ Y_{i=2} \\ \vdots \\ Y_{i=n} \end{bmatrix} = \begin{bmatrix} 1 & X_{i=1} \\ 1 & X_{i=2} \\ \vdots & \vdots \\ 1 & X_{i=n} \end{bmatrix} \begin{bmatrix} b_o \\ b_1 \end{bmatrix} + \begin{bmatrix} e_{i=1} \\ e_{i=2} \\ \vdots \\ e_{i=n} \end{bmatrix}$$

where b_o = intercept and
b_1 = slope.

Finally,

$$\begin{bmatrix} Y_{i=1} \\ Y_{i=2} \\ \vdots \\ Y_{i=n} \end{bmatrix} = \begin{bmatrix} b_o + b_1X_{i=1} \\ b_o + b_1X_{i=2} \\ \vdots \\ b_o + b_1X_{i=n} \end{bmatrix} + \begin{bmatrix} e_{i=1} \\ e_{i=2} \\ \vdots \\ e_{i=n} \end{bmatrix}$$

$$= \begin{bmatrix} b_o + b_1X_{i=1} + e_{i=1} \\ b_o + b_1X_{i=2} + e_{i=2} \\ \vdots \\ b_o + b_1X_{i=n} + e_{i=n} \end{bmatrix}$$

We want to find B such that e'e is a minimum, where

$$e'e = \begin{bmatrix} e_{i=1} & e_{i=2} & \text{----} & e_{i=n} \end{bmatrix} \begin{bmatrix} e_{i=1} \\ e_{i=2} \\ \vdots \\ e_{i=n} \end{bmatrix} = \sum_i e_i^2 \quad .$$

This is the "least squares solution".

If
$$\begin{aligned} e &= Y - XB, \quad \text{then} \\ e'e &= (Y-XB)'(Y-XB) \\ &= Y'Y - (XB)'Y - Y'XB + (XB)'XB \\ &= Y'Y - 2X'YB + X'XB'B, \\ &\qquad\qquad \text{because } X'Y = Y'X \end{aligned}$$

To minimize something, we differentiate it with respect to the parameter(s) we are trying to estimate and set it equal to zero,

so
$$\frac{\partial e'e}{\partial B} = 0 = -2X'Y + 2X'XB,$$

and
$$X'XB = X'Y,$$
and
$$\hat{B} = (X'X)^{-1} X'Y \ .$$

In subscripted, rather than matrix, notation

$$X'XB = X'Y \qquad \text{is}$$

$$\begin{bmatrix} 1 & 1 & \text{------} & 1 \\ X_{i=1} & X_{i=2} & \text{---} & X_{i=n} \end{bmatrix} \begin{bmatrix} 1 & X_{i=1} \\ 1 & X_{i=2} \\ \vdots & \vdots \\ 1 & X_{i=n} \end{bmatrix} \begin{bmatrix} b_o \\ b_1 \end{bmatrix} \quad \begin{bmatrix} 1 & 1 & \text{----} & 1 \\ X_{i=1} & X_{i=2} & \text{----} & X_{i=n} \end{bmatrix} \begin{bmatrix} Y_{i=1} \\ Y_{i=2} \\ \vdots \\ Y_{i=n} \end{bmatrix}$$

$$\begin{bmatrix} n & \sum_i X_i \\ \sum_i X_i & \sum_i X_i^2 \end{bmatrix} \begin{bmatrix} b_o \\ b_1 \end{bmatrix} = \begin{bmatrix} \sum_i Y_i \\ \sum_i X_i Y_i \end{bmatrix}$$

$$b_o n + b_1 \sum_i X_i = \sum_i Y_i$$

$$b_o \sum_i X_i + b_1 \sum_i X_i^2 = \sum_i X_i Y_i$$

$$-b_o \sum_i X_i - b_1 \sum_i X_i (\sum_i X_i/n) = - \sum_i Y_i (\sum_i X_i/n)$$

$$b_o \sum_i X_i + b_1 \sum_i X_i^2 = \sum_i X_i Y_i$$

$$b_1 \left(\sum_i X_i^2 - \frac{(\sum_i X_i)^2}{n} \right) = \sum_i X_i Y_i - \frac{\sum_i X_i \sum_i Y_i}{n}$$

$$\hat{b}_1 = \frac{\sum xy}{\sum x} .$$

Then

$$b_o = (\sum_i Y_i - b_1 \sum_i X_i)/n$$

$$= \frac{\sum_i Y_i}{n} - b_1 \frac{\sum_i X_i}{n} \quad \text{and}$$

$$\hat{b}_o = \bar{Y} - b_1 \bar{X} .$$

3.5 Practical session 1

3.5.1 Assignment/Tutorial

Purpose:

(1) To provide practice in using the APPLEs, the IBM PC, and the IBM mainframe.

(2) To provide practice in running BASIC, FORTRAN and APL programs, and in using the MINITAB statistical package.

(3) To provide some review in simple regression analysis.

Approach:

Given a small set of observations on variables X and Y:

X:	1	2	3	4
Y:	2	4	6	7

use the above-mentioned hardware/software to

(1) plot Y versus X

(2) calculate the least squares regression of Y on X.

Procedure:

(1) On the APPLE, use DOS 3.3 to run the programs REGRESSION and PLOT. These programs require that the X,Y data are in a sequential text file, which can be created using the program CREATE TEXT. All 3 of these BASIC programs are from Orloci and Kenkel (1984).

(2) On the APPLE, use CP/M to run the program LINREG. You can choose the option of entering data from the screen, or the option of reading data from a file. If you choose the 2nd option, then you first have to run program CREATE, which will create a data file called 'DATA'.

(3) On the computer terminal to the IBM mainframe (including IBM PCs used as terminals), run the program LINREG. When you are in BASIC mode, enter 'LOAD LINREG'. Then enter 'RUN', and select the option to enter data from the keyboard. Enter the data as '1,2', '2,4', '3,6', and '4,7'. The t-value for 2 error df is 4.3. When the program offers more statistics, respond with a 'Y'. (N.B.: All letter responses must be capital letters.) After you have run the program, you can enter 'LIST' and if you have some knowledge of BASIC you can examine the program to see how it is designed.

(4) On the computer terminal to the IBM mainframe, run the FORTRAN programs PLOT and REGR. But first, after you have logged on and done your 'DEF STOR 1500K', enter 'TYPE REGR DATA'. Enter 'COPY REGR DATA A PLOT DATA A' to produce an identical data set for the PLOT program. Note the last line which indicates "end of file". Now enter 'TYPE PLOT FORTRAN'. It is rather long and tedious - a utility program after all - but you should look at the first screenfull to see how the data are read in from the file 'PLOT DATA'. Then enter 'TYPE REGR FORTRAN', and do the same. Now run the program 'PLOT FORTRAN' by

entering 'FORTVS PLOT'. (The name 'FORTVS' calls an 'EXEC' file which has been set up to do the compiling, loading, running, and to identify the correct input and output files. If you are interested, you can look at this 'EXEC' file by entering 'TYPE FORTVS EXEC'.) The information that appears on your screen is diagnostics of the progress of what 'FORTVS EXEC' is doing. When you see ';R------', then the program has run. Your output is in file 'PLOT OUTPUT'. To see it, enter 'TYPE PLOT OUTPUT'. Some parameters of the program run are shown, and then the plot, which may be split between screenfuls. To cure this, go into the editor (type 'XEDIT PLOT OUTPUT') and then delete all the lines except the plot itself. This can be done by putting 'DD' on the dashed "prefix" lines: put 'DD' on the first line of the file and put 'DD' on the last line before the plot itself. Then depress ENTER and all the lines before the plot will disappear. Enter 'FILE', and then clear the screen and enter 'TYPE PLOT OUTPUT' again.

Now run 'REGR FORTRAN' in the same manner. When it has run, enter 'TYPE REGR OUTPUT'.
If you want hard copy of your data files and/or your output files, they can be printed out using 'PRINT' or 'LPRINT'.

In running FORTRAN programs you automatically create files. The compiler creates a machine language file 'fn TEXT', and the run creates a file 'fn LISTING' with run-time diagnostics in it (which you should look at if the program didn't work). And of course there is 'fn OUTPUT' created as well. Before logging off you should always erase such files if you have no use for them. If you don't, then in a few days you will have your disk area filled with "junk files".

(5) On the IBM PC, run the BASIC program LINREG. (This is the same BASIC program you ran on the IBM mainframe.)
The program is on the "MS DOS program" disk.
Again, choose the option to enter data from the keyboard.
The FORTRAN programs PLOT and REGR can also be run on the IBM PC.
(6) On the computer terminal to the IBM mainframe, run the APL

"programs" (called functions in APL) SCATTERPLOT and GLM. Once in APL mode, proceed as follows.

```
)LOAD UNESCO           This loads the workspace.

)FNS                   This causes the functions in  this
                         workspace to be listed.

)VARS                  This causes the variables in  this
                         workspace  to  be  listed.  (There
                         aren't any initially.)

X← 1 2 3 4             This creates a variable  X  which
                         contains the vector of numbers '1 2 3 4'.

X                      This displays the contents of X on
                         the screen.

Y← 2 4 6 7             These two commands do the same
Y                        for variable Y.

         (now clear the screen)

X SCATTERPLOT Y        This "runs" the SCATTERPLOT function.

34 20                  is your response (as suggested).

N                      is your response - you do not want
                         to force plot axes to have the same
                         scale.  Follow instructions.

X← 4 1⍴X               These two commands change variable
X                        X from a vector to a 4-by-1 matrix,
                         and then display it on the screen.

Y← 4 1⍴Y               Same for
Y                        variable Y.
D← X,[2]Y              These two commands catenate X and
```

```
D                            Y into a new variable D, and
                               display it on the screen.

        (now clear the screen)

1 2 GLM D                    This "runs" the GLM function.  The
                               '1 2' says that the independent (X)
                               variable is in column 1 and the
                               dependent (Y) variable is in column 2.
                               The variable D contains the data.  You
                               can have more than 1 X-variable
                               (multiple regression).   The  first
                               column  at  the bottom left is  the
                               predicted Y values.   The second is
                               the Y residuals.

)OFF HOLD                    Leaves APL mode but keeps you
                               logged on to the system.
```

You can also do this exercise on the IBM PC which has APL implemented on it.

(7) On the IBM mainframe, do the same plot and regression analysis in MINITAB, as follows.

```
READ INTO C1-C2
1 2
2 4                          Reads the data into C1 and C2 and
3 6                            then prints the contents of C1 and C2.
4 7
PRINT C1-C2

WIDTH 55, HEIGHT 16          Changes plot dimensions to fit the
(clear the screen)             screen, then after clearing the
PLOT C2 VS C1                  screen, plots Y versus X.


PLOT C2 FROM 0 TO 10           The same plot, but with you (rather
  VS C1 FROM 0 TO 5         than MINITAB) controlling the scales
                            on the axes.

REGRESS C2 ON 1 PRED.     Does the regression of
  IN C1                     Y on X.

BRIEF 1                   Limits the regression output, repeats
REGR C2 1 C1, C3, C4      the regression storing Y residuals in C3
PRINT C3-C4               and Y predicted in C4,
PLOT C4 VS C1             prints the contents of C3 and C4
PLOT C3 VS C4             plots the fitted line, and
                          plots Y residuals versus Y predicteds.

STOP                      Leaves MINITAB but stays on the system.
```

3.6 PRACTICAL SESSION 2

3.6.1 Assignment/Tutorial

3.6.1.1 MINITAB

A. READ in the data set from file 'REGR1 DATA' into C1 and C2. There are 100 observations by 2 variables (Y and X respectively).

B. Do a Model I regression analysis by matrix algebra:

1. PLOT Y versus X to check that a linear regression model looks sensible.
2. SET 100 values of 1 into C3, and then COPY C3 and C2 into M1. The X matrix is now in M1.
3. TRANSPOSE M1 and put it into M2. The X' matrix is now in M2.
4. MULTIPLY M2 by M1 and put the product X'X into M3.
5. COPY C1 into M4. The Y matrix is now in M4.
6. MULTIPLY M2 by M4 and put the product X'Y into M5.
7. INVERT M3 and put X'X inverse into M6.

8. MULTIPLY M6 by M5 and put the product, (X'X inverse)*(X'Y), into M7. The matrix of regression coefficients, B, is now in M7.
9. MULTIPLY M1 by M7 and put the product XB into M8. The predicted Y values (Y-hat) are now in M8.
10. COPY M8 into C4, and then LET K1=SUM((C1-C4)*(C1-C4)). The Error SS is now in K1.
11. LET K2=SUM((C1-AVER(C1))*(C1-AVER(C1))). The total SS is now in K2.
12. LET K3=K2-K1 puts the Regression SS into K3. LET K4=K3/K2 puts r-squared into K4. LET K5=K1/98 puts the Error MS into K5. LET K6=K3/K5 puts F into K6. You have your ANOVA table.

C. Now do the same analysis using the REGR command:

1. Do BRIEF 1. Then do REGR C1 1 C2, and then BRIEF 6 followed by REGR C1 1 C2. Compare these outputs with each other and with the results from the matrix algebra solution.
2. Do BRIEF 1 and then REGR C1 1 C2,C5,C6. This time you have put the standardized residuals, (Y - Y-hat)/SQRT(Error MS), and the predicted Y values, Y-hat, into C5 and C6 respectively. Compare them with your BRIEF 6 output. (To convert the standardized residuals back to the "raw" residuals you would just multiply C5 by SQRT(Error MS)=SQRT(K5).)
3. Do HISTOGRAM of C5 to see if the residual errors e = Y - Y-hat appear to be normally distributed. Another way to to it is by an arithmetic probability plot, by doing NSCORES C5,C7 and then PLOT C5 C7. If the residual errors are approximately normal, you should see a fairly straight line.
4. Now see whether the residual errors are independent of the other effects in the model, as they should be. PLOT C5 versus C6. You should see a normally distributed scatter centred on e=C5=0. There should

be no pattern, or relationships, apparent in the plot.

5. Do RLINE C1,C2 which produces estimates of slope and intercept by an iterative procedure quite different from the least-squares estimate and more robust to outliers. Compare these estimates with the least-squares estimates.

3.6.1.2 SAS

We will run SAS in batch mode (although it can be run in interactive mode). First, we must create a file containing the SAS job commands. Start by entering 'XEDIT RUNSAS SAS', and then go into INPUT mode within the editor. Enter the following lines:

DATA REGSAS;	Names the SAS data set to be created.
INPUT Y X;	Names the variables and their input order.
CARDS;	Says the data follow, on "card images".
PROC PRINT;	Causes the data just read in to be printed.
PROC PLOT; PLOT Y*X;	Produces a plot of Y versus X.
PROC GLM; MODEL Y=X;	Produces a regression analysis of Y on X.

Note that all SAS statements end with a ';'. You can have several statements to a line, or a statement can continue over several lines, as long as you remember to end each statement with a semi-colon.

Now, what about the data? A SAS job can read data from a data file, but it is easier to just "pull" the data into the SAS job file we have just created. Get out of INPUT mode. Move the "active line" (the brightly lit up line) up or down until the 'CARDS;' line is the active line. (Use the commands 'UP' or 'DOWN' to move the active line - for example if the bottom line is the active line then 'UP 3' should do it.) Now enter 'GET REGR1 DATA', and all the data should be inserted into the file after the 'CARDS;' line. Now enter 'FILE' to leave the editor, and then run your SAS job by entering 'SAS RUNSAS'. When you get the response ';R----', the job has run. Enter 'TYPE RUNSAS SASLOG' to see a record of the run. To see the output, enter 'TYPE RUNSAS LISTING'. Notice that the output is intended for printing on 132-character-wide paper, not for display on an 80-character-wide screen. To print out hard copy on the local printer, enter 'LPRINT B08 fn ft'. I would suggest that you print out both 'RUNSAS SAS' (the job command file) and 'RUNSAS LISTING' (the output file).

3.6.2 Job Listings and Outputs.

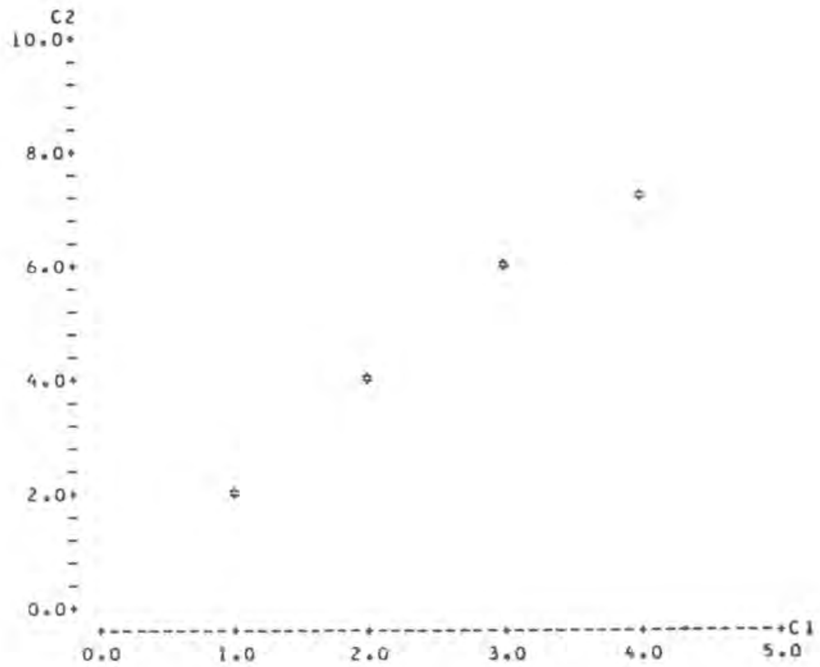
```
FILE: REGR      MINITAB  A   VM/SP - CONVERSATIONAL MONITOR SYSTEM                              PAGE 001


READ INTO C1-C2
1 2
2 4
3 6
4 7
PRINT C1-C2
PLOT C2 VS C1
PLOT C2 FROM 0 TC 10 VS C1 FROM 0 TO 5
REGRESS C2 ON 1 PRED. IN C1
BRIEF 1
REGR C2 1 C1, C3, C4
PRINT C3-C4
PLOT C4 VS C1
PLOT C3 VS C4
STOP
```

MINITAB Job Listing for regression run on small data set.

MINITAB output from regression run on small data set.

```
FILE: REGR      OUTPUT   A   VM/SP - CONVERSATIONAL MONITOR SYSTEM                              PAGE 001

1MINITAB RELEASE 81.1 *** COPYRIGHT - PENN STATE UNIV. 1981
 MAY 22, 1985 *** NATIONAL UNIVERSITY OF SINGAPORE - LOCAL VERSION 02/12/1982
 STORAGE AVAILABLE    4800

--
--
--
--
--
--
 COLUMN        C1          C2
 COUNT          4           4
 ROW
   1           1.          2.
   2           2.          4.
   3           3.          6.
   4           4.          7.

--
              C2
             7.0+                                         *
                -
                -
                -
                -
             6.0+                             *
                -
                -
                -
                -
             5.0+
                -
                -
                -
                -
             4.0+                *
                -
                -
                -
             3.0+
                -
                -
                -
             2.0+    *
                  +---------+---------+---------+---------+---------+C1
                0.80      1.60      2.40      3.20      4.00      4.80
```

```
FILE: REGR      OUTPUT   A   VM/SP - CONVERSATIONAL MONITOR SYSTEM                              PAGE 002

--
```



```
--

THE REGRESSION EQUATION IS
Y =    0.500 +  1.70 X1

                                          ST. DEV.     T-RATIO =
        COLUMN       COEFFICIENT          OF COEF.     COEF/S.D.
          --             0.5000            0.4743         1.05
X1      C1               1.7000            0.1732         9.81

THE ST. DEV. OF Y ABOUT REGRESSION LINE IS
S = 0.3873
WITH (   4- 2) =   2 DEGREES OF FREEDOM

R-SQUARED = 98.0 PERCENT
R-SQUARED = 96.7 PERCENT, ADJUSTED FOR D.F.

ANALYSIS OF VARIANCE

 DUE TO      DF            SS       MS=SS/DF
```

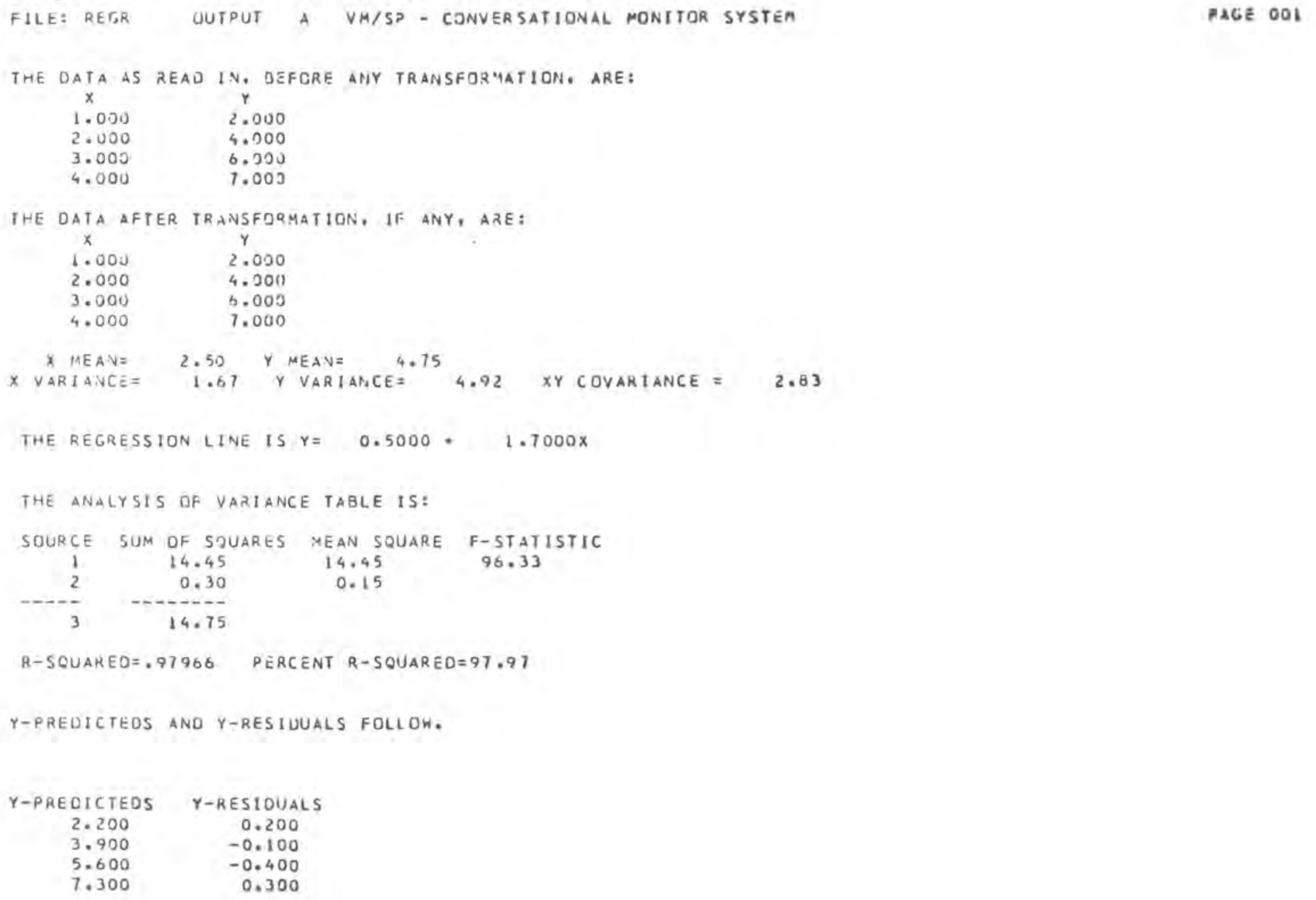
```
FILE: REGR     OUTPUT   A   VM/SP - CONVERSATIONAL MONITOR SYSTEM                          PAGE 003

 REGRESSION   1      14.4500      14.4500
 RESIDUAL     2       0.3000       0.1500
 TOTAL        3      14.7500


 DURBIN-WATSON STATISTIC = 2.23
 --
 --

                                        ST. DEV.     T-RATIO =
        COLUMN      COEFFICIENT         OF COEF.     COEF/S.D.
         --              0.5000           0.4743          1.05
 X1      C1              1.7000           0.1732          9.81

 S = 0.3873

 R-SQUARED = 98.0 PERCENT
 --
 COLUMN         C3          C4
 COUNT           4           4
 ROW
   1     -0.94281     2.20000
   2      0.30861     3.90000
   3      1.23443     5.60000
   4     -1.41422     7.30000

 --
            C4
         7.6+
            -                                     *
            -
            -
            -
         6.2+
            -
            -                        *
            -
            -
         4.8+
            -
            -
            -            *
            -
         3.4+
            -
            -
            -
            -   *
         2.0+
              +---------+---------+---------+---------+---------+C1
            0.80      1.60      2.40      3.20      4.00      4.80
```

LINREG on small data set on PC.

```
Ok
run
HOW MANY X-Y PAIRS DO YOU WANT TO ENTER?
? 4

IF YOU WANT NO TRANSFORMATION OF X INPUT 0
IF YOU WANT A LOG(X) TRANSFORMATION INPUT 1
IF YOU WANT A LOG(X+1) TRANSFORMATION INPUT 2
? 0
IF YOU WANT NO TRANSFORMATION OF Y INPUT 0
IF YOU WANT A LOG(Y) TRANSFORMATION INPUT 1
IF YOU WANT A LOG(Y+1) TRANSFORMATION INPUT 2
? 0

ENTER THE DATA AS X-Y PAIRS.

? 1,2
? 2,4
? 3,6
? 4,7

WHAT IS THE T-VALUE FOR THE 95% CONFIDENCE LIMITS
WITH   2   DEGREES OF FREEDOM?
? 4.303


WHAT IS THE T-VALUE FOR THE 95% CONFIDENCE LIMITS
WITH   2   DEGREES OF FREEDOM?
? 4.303


THE REGRESSION STATISTICS ARE AS FOLLOWS:

THE EQUATION OF THE LINE IS:  Y= .5 + 1.7 X

WHERE THE SLOPE IS:   1.7
AND THE Y-INTERCEPT IS:   .5                 2
THE STANDARD ERROR OF THE REGRESSION IS:  (+ OR -)  5.960285
THE STANDARD ERROR OF THE SLOPE IS:  (+ OR -)  2.665521
THE 95% C.L. FOR THE SLOPE ARE:  -9.769736       13.16974
THE STANDARD ERROR OF THE INTERCEPT IS:  (+ OR -)  7.299829
THE 95% C.L. FOR THE INTERCEPT ARE:  -30.91116       31.91116
THE CORRELATION COEFFICIENT (R) IS:   .4111032
THE COEFFICIENT OF DETERMINATION (R**2) IS:   .1690058


DO YOU WANT MORE STATISTICS PRINTED?
TYPE Y OR N.
? Y
```

```
THE REGRESSION COMPUTATIONS HAVE PRODUCED THE FOLLOWING:

THE MEANS OF X AND Y ARE:   2.5       4.75
THE SUM OF X IS:   10
THE SUM OF Y IS:   19
THE SUM OF X-SQUARED IS:   30
THE SUM OF Y-SQUARED IS:   105
THE SUM OF X*Y IS:   56
THE SUM OF SQUARES OF X IS:   5
THE SUM OF CROSS-PRODUCTS IS:   8.5
THE REGRESSION SUM OF SQUARES IS:   14.45
THE RESIDUAL SUM OF SQUARES IS:   71.05
THE TOTAL SUM OF SQUARES IS:   85.5
THE REGRESSION MEAN SQUARE IS:   14.45
THE RESIDUAL MEAN SQUARE IS:   35.525

THE F-VALUE IS:   .4067558
WITH 1 REGRESSION D OF F, AND
A RESIDUAL D OF F OF :   2


DO YOU WANT 95% CONFIDENCE LIMITS?
TYPE Y OR N.
? Y


THE RESIDUAL MEAN SQUARE IS:   35.525

THE F-VALUE IS:   .4067558
WITH 1 REGRESSION D OF F, AND
A RESIDUAL D OF F OF :   2


DO YOU WANT 95% CONFIDENCE LIMITS?
TYPE Y OR N.
? Y

DO YOU WANT TO SPECIFY THE X VALUES?
TYPE Y OR N.
? N

THE PREDICTED VALUES AND 95% C.L. OF Y ARE:

GIVEN X VALUE   PREDICTED Y     LOWER Y        UPPER Y       ERROR
   1     2.2   -19.25791     23.65791     4.986733
   2     3.9   -10.1475     17.9475     3.264583
   3     5.600001    -8.447499     19.6475     3.264583
   4     7.3   -14.15791     28.75791     4.986733
Ok
```

Output of run of REGR. FORTRAN on small data set.



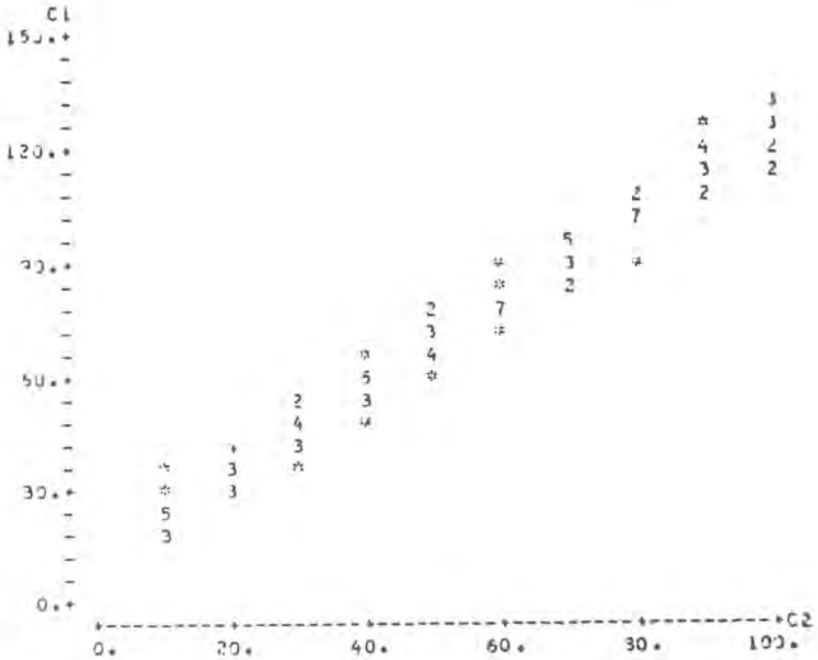
```
FILE: REGR      OUTPUT   A   VM/SP - CONVERSATIONAL MONITOR SYSTEM                                   PAGE 001

THE DATA AS READ IN, BEFORE ANY TRANSFORMATION, ARE:
      X            Y
    1.000        2.000
    2.000        4.000
    3.000        6.000
    4.000        7.000

THE DATA AFTER TRANSFORMATION, IF ANY, ARE:
      X            Y
    1.000        2.000
    2.000        4.000
    3.000        6.000
    4.000        7.000

   X MEAN=     2.50   Y MEAN=     4.75
X VARIANCE=     1.67   Y VARIANCE=     4.92   XY COVARIANCE =     2.83


 THE REGRESSION LINE IS Y=   0.5000 +    1.7000X


 THE ANALYSIS OF VARIANCE TABLE IS:

 SOURCE  SUM OF SQUARES  MEAN SQUARE  F-STATISTIC
     1        14.45         14.45        96.33
     2         0.30          0.15
 -----    --------
     3        14.75

 R-SQUARED=.97966   PERCENT R-SQUARED=97.97


Y-PREDICTEDS AND Y-RESIDUALS FOLLOW.


Y-PREDICTEDS   Y-RESIDUALS
     2.200        0.200
     3.900       -0.100
     5.600       -0.400
     7.300        0.300
```

```
NOTE THIS IS A LINEAR REGRESSION ANALYSIS BY MINITAB
READ C1-C2
    19.41    10.
    23.15    10.
    33.62    10.
    26.72    10.
    24.60    10.
    23.50    10.
    26.59    10.
    29.83    10.
    19.37    10.
    20.31    10.
    42.93    20.
    40.16    20.
    35.11    20.
    41.14    20.
    33.52    20.
    31.93    20.
    39.58    20.
    32.69    20.
    37.65    20.
    31.34    20.
    44.53    30.
    52.15    30.
    47.24    30.
    46.35    30.
    48.71    30.
    43.38    30.
    52.15    30.
    46.46    30.
    44.42    30.
    37.19    30.
    60.67    40.
    61.03    40.
    52.24    40.
    55.46    40.
    59.02    40.
    49.93    40.
    53.08    40.
    58.92    40.
    57.55    40.
    61.93    40.
    59.96    50.
    76.33    50.
    77.54    50.
    63.40    50.
    74.75    50.
    74.16    50.
    66.59    50.
    68.73    50.
    62.55    50.
    53.30    50.
    78.70    60.
    77.97    60.
    79.02    60.
```

```
     78.51    60.
     81.95    60.
     90.25    60.
     77.24    60.
     73.85    60.
     77.75    60.
     78.00    60.
     86.80    70.
     97.50    70.
     98.95    70.
     84.13    70.
     91.24    70.
     96.95    70.
     95.62    70.
     89.95    70.
     91.47    70.
     95.85    70.
    104.56    80.
    105.35    80.
    101.84    80.
    104.96    80.
    103.10    80.
    110.26    80.
    101.77    80.
    101.02    80.
    100.47    80.
     91.80    80.
    117.65    90.
    112.62    90.
    115.45    90.
    106.53    90.
    119.42    90.
    114.69    90.
    125.96    90.
    122.10    90.
    110.82    90.
    119.41    90.
    131.95   100.
    126.28   100.
    132.22   100.
    111.13   100.
    129.10   100.
    125.25   100.
    120.95   100.
    112.01   100.
    128.53   100.
    112.11   100.
  PRINT C1-C2
  PLOT C1 VS C2
  SET C3
  100(1)
  COPY C3 AND C2 INTO M1
  TRANSPOSE M1, PUT IN M2
  MULTIPLY M2 BY M1, PUT PRODUCT IN M3
  PRINT M3
```

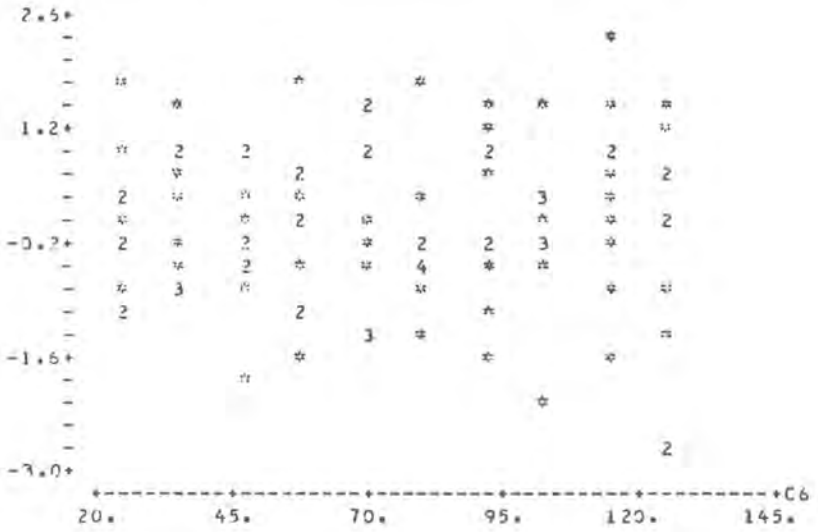
```
COPY C1 INTO M4
NOTE M3 IS THE X'X MATRIX
MULTIPLY M2 BY M4, PUT PRODUCT IN M5
PRINT M5
NOTE M5 IS THE X'Y MATRIX
INVERT M3, PUT IN M6
MULTIPLY M6 BY M5, PUT PRODUCT IN M7
PRINT M7
NOTE M7 IS THE B MATRIX
MULTIPLY M1 BY M7, PUT PRODUCT IN M8
COPY M8 INTO C4
LET K1=SUM((C1-C4)*(C1-C4))
LET K2=SUM((C1-AVER(C1))*(C1-AVER(C1)))
LET K3=K2-K1
LET K4=K3/K2
LET K5=K1/98
LET K6=K3/K5
PRINT K2
NOTE K2 IS THE TOTAL SUM OF SQUARED DEVIATIONS
PRINT K1
NOTE K1 IS THE ERROR SUM OF SQUARED DEVIATIONS
PRINT K3
NOTE K3 IS THE REGRESSION SUM OF SQUARED DEVIATIONS
PRINT K3
NOTE K3 IS ALSO THE REGRESSION MEAN SQUARED DEVIATIONS
PRINT K5
NOTE K5 IS THE ERROR MEAN SQUARED DEVIATIONS
PRINT K6
NOTE K6 IS THE F RATIO
PRINT K4
NOTE K4 IS THE COEFFICIENT OF DETERMINATION R-SQUARED

NOTE TO RUN LINEAR REGRESSION USING THE REGR FUNCTION
BRIEF 6
REGRESS C1 1 C2, C5, C6
HIST C5
SQRT(K5), K7
MULTIPLY C5 BY K7, C8
PRINT C8
NSCORES C5, C7
PLOT C5 VS C6
PLOT C5 VS C7
RLINE C1,C2
STOP
```

```
FILE: LAB2A    OUTPUT   A1  VM/SP - CONVERSATIONAL MONITOR SYSTEM                                    PAGE 001

1MINITAB RELEASE 81.1 *** COPYRIGHT - PENN STATE UNIV. 1981
 APRIL 29, 1985 *** NATIONAL UNIVERSITY OF SINGAPORE - LOCAL VERSION 02/12/1982
 EXAMPLE OF A LINEAR REGRESSION ANALYSIS USING MINITAB
 --
 COLUMN        C1            C2
 COUNT        100           100
 ROW
   1       19.410           10.
   2       23.150           10.
   3       33.620           10.
   4       25.720           10.
   5       24.600           10.
   6       23.500           10.
   7       26.590           10.
   8       29.330           10.
   9       17.370           10.
  10       20.310           10.
  11       42.930           20.
  12       40.160           20.
  13       35.110           20.
  14       41.140           20.
  15       33.520           20.
  16       31.970           20.
  17       39.590           20.
  18       32.690           20.
  19       37.550           20.
  20       31.340           20.
  21       44.530           30.
  22       52.150           30.
  23       47.240           30.
  24       46.050           30.
  25       43.710           30.
  26       43.390           30.
  27       52.150           30.
  28       46.460           30.
  29       44.420           30.
  30       37.170           30.
  31       60.570           40.
  32       61.030           40.
  33       52.240           40.
  34       55.490           40.
  35       59.020           40.
  36       49.730           40.
  37       53.080           40.
  38       58.920           40.
  39       67.550           40.
  40       61.930           40.
  41       67.960           50.
  42       76.330           50.
  43       77.340           50.
  44       63.400           50.
  45       74.750           50.
  46       74.160           50.
  47       66.590           50.
  48       69.730           50.
  49      62.550       50.
  50      63.300       50.
  51      78.700       60.
  52      77.270       60.
  53      79.320       60.
  54      73.510       60.
  55      81.350       60.
  56      90.250       60.
  57      77.240       60.
  58      73.350       60.
  59      79.750       60.
  60      73.000       60.
  61      86.300       70.
  62      97.500       70.
  63      93.950       70.
  64      84.130       70.
  65      91.240       70.
  66      96.950       70.
  67      95.620       70.
  68      99.330       70.
  69      91.470       70.
  70      95.350       70.
  71     104.560       80.
  72     105.350       80.
  73     101.840       80.
  74     104.960       80.
  75     103.100       80.
  76     110.250       80.
  77     101.770       80.
  78     101.890       80.
  79     100.470       80.
  80      91.300       80.
  81     117.550       90.
  82     112.620       90.
  83     115.450       90.
  84     106.530       90.
  85     119.420       90.
  86     114.590       90.
  87     125.960       90.
  88     122.100       90.
  89     110.820       90.
  90     119.410       90.
  91     131.950      100.
  92     126.290      100.
  93     132.220      100.
  94     111.130      100.
  95     127.100      100.
  96     125.250      100.
  97     120.950      100.
  98     119.010      100.
  99     123.580      100.
 100     112.110      100.

--
 PLOT OF Y AGAINST X
```

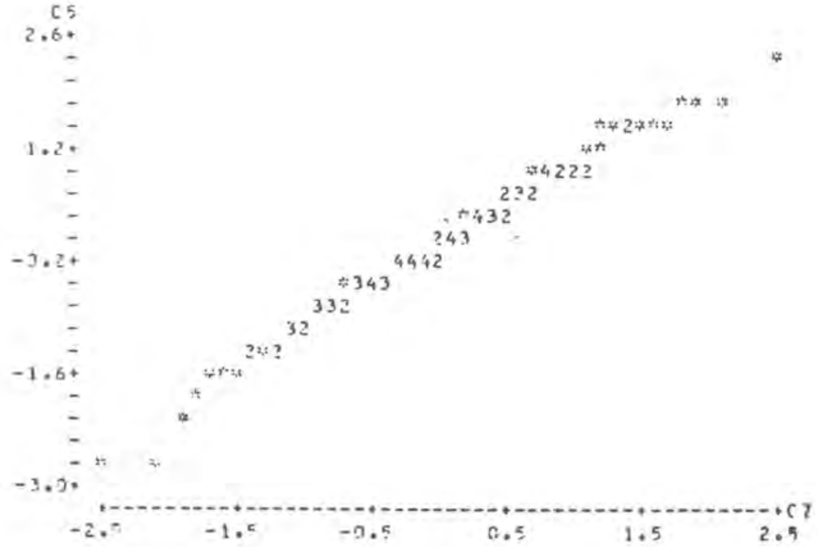


```
--
--
MATRIX M3 IS THE X'X MATRIX
--
   MATRIX M3                  2 ROWS BY     2 COLUMNS

         100.       5500.
        5500.     335000.

--
 MATRIX M5 IS THE X'Y MATRIX
--
   MATRIX M5                  2 ROWS BY     1 COLUMNS

         7504.
       505005.

--
 MATRIX M7 IS THE B MATRIX
--
   MATRIX M7                  2 ROWS BY     1 COLUMNS

      13.5161
       1.1196
```

```
FILE: LA02A    OUTPUT    A1  VM/SP - CONVERSATIONAL MONITOR SYSTEM                                        PAGE 004

  K2 IS THE TOTAL SUM OF SQUARED DEVIATIONS

   K2        105827.
--
  K1 IS THE RESIDUAL SUM OF SQUARED DEVIATIONS
--
   K1        2595.04
--
  K3 IS THE REGRESSION SUM OF SQUARED DEVIATIONS
--
   K3        103231.
--
  K3 IS ALSO THE REGRESSION MEAN SQUARED DEVIATIONS
--
   K3        103231.
--
  K5 IS THE ERROR MEAN SQUARED DEVIATIONS
--
   K5        26.4702
--
  K6 IS THE F RATIO
--
   K6        3826.95
--
  K4 IS THE COEFFICIENT OF DETERMINATION, R-SQUARED
--
   K4       0.975469
--

 THE REGRESSION EQUATION IS
 Y =    13.5 +  1.12 X1

                                              ST. DEV.     T-RATIO =
        COLUMN      COEFFICIENT               OF COEF.     COEF/S.D.
        INTERCEPT      13.519                  1.112         12.15
        SLOPE          1.11859                0.01792        62.43

 THE ST. DEV. OF Y ABOUT REGRESSION LINE IS
 S = 5.147
 WITH ( 100- 2) =  98 DEGREES OF FREEDOM

 R-SQUARED = 97.5 PERCENT
 R-SQUARED = 97.5 PERCENT, ADJUSTED FOR D.F.

 ANALYSIS OF VARIANCE

  DUE TO       DF            SS         MS=SS/DF
 REGRESSION   1       103229.6         103229.6
 RESIDUAL     98        2595.0             26.5
 TOTAL        99      105825.7

              X1              Y     PRED. Y      ST.DEV.
 ROW          C2             C1       VALUE      PRED. Y     RESIDUAL       ST.RES.
   1          10         19.410      24.704        0.957       -5.294         -1.05
 (X-PRIME X)INVERSE

                0            1
   0    0.0466645
   1   -0.0006606    0.0000121

--

  HISTOGRAM OF C5, THE RESIDUALS

   MIDDLE OF    NUMBER OF
   INTERVAL     OBSERVATIONS
      -3.0         1    *
      -2.5         1    *
      -2.0         2    **
      -1.5         6    ******
      -1.0        11    ***********
      -0.5        18    ******************
       0.0        22    **********************
       0.5        15    ***************
       1.0        12    ************
       1.5         8    ********
       2.0         3    ***
       2.5         1    *

--
   ANSWER =          5.1469
--
--
 COLUMN         C3
 COUNT         100
      -5.3876     -1.5912      0.0745      2.0521     -0.1055     -1.2250
       1.9198      5.2173     -5.4283     -4.4716      7.1296      4.3245
      -0.7834      5.3170     -2.3976     -4.0097      3.7372     -3.2401
       1.7827     -4.6072     -2.5177      5.1198      0.1660     -1.0345
       1.6471     -3.7283      5.1199     -0.6209     -2.6791     -9.9735
       2.4242      2.7355     -6.0600     -2.7992      0.7535     -8.3949
      -5.2145      0.6629      9.3484      3.6923      0.5155      7.4212
       8.1340     -5.0736      5.3304      4.7373     -2.8720     -0.7209
      -6.9330     -5.1791     -1.7431     -3.6769     -1.6216     -2.1341
       1.2233      9.6659     -3.4107     -5.8183     -0.8976     -2.5467
      -5.0512      5.7176      7.1769     -7.7334     -0.5926      5.1741
       3.3255     -1.8507     -0.3512      4.0570      1.5591      2.3661
      -1.1752      1.9726      0.0930      7.3198     -1.2459     -1.1343
      -2.5574    -11.3046      3.5031     -1.5906      1.2753     -7.7577
       5.2956      0.5057     11.9154      3.0095     -3.4134      5.2954
       6.6901      0.9195      5.9648    -14.4993      3.7895     -0.1289
      -7.5051     -6.4795      3.2603    -13.5019

1
--
   PLOT OF RESIDUALS AGAINST PREDICTED Y
--
               C5
```



```
PLOT OF C5 AGAINST C7
--
```



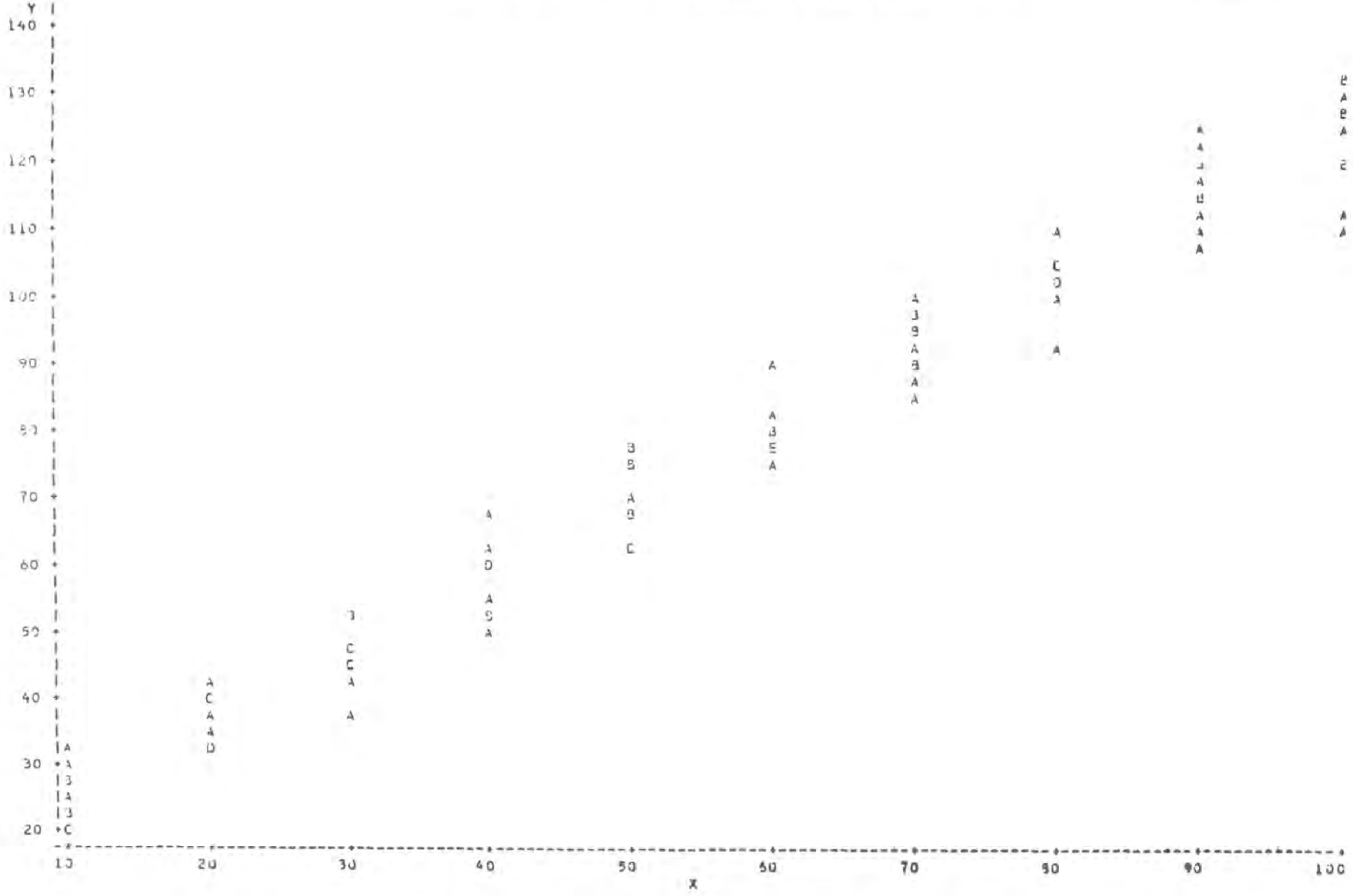
```
ESTIMATES OF INTERCEPT AND SLOPE BY ITERATIVE METHOD
```

```
FILE: RUNSAS    SAS       A1  VM/SP - CONVERSATIONAL MONITOR SYSTEM                                PAGE 001

TITLE RUNNING LINEAR REGRESSION USING PROC GLM ON SAS;
DATA REGSAS;
INPUT Y X;
CARDS;
    17.41    10.
    23.15    10.
    33.62    10.
    26.72    10.
    24.60    10.
    23.50    10.
    25.59    10.
    29.83    10.
    19.37    10.
    20.31    10.
    42.93    20.
    40.16    20.
    35.11    20.
    41.14    20.
    33.52    20.
    31.93    20.
    39.58    20.
    32.69    20.
    37.65    20.
    31.34    20.
    44.59    30.
    52.15    30.
    47.24    30.
    45.05    30.
    43.71    30.
    43.33    30.
    52.15    30.
    46.46    30.
    44.42    30.
    37.19    30.
    60.67    40.
    61.03    40.
    52.24    40.
    55.43    40.
    59.02    40.
    49.93    40.
    53.05    40.
    58.92    40.
    67.55    40.
    51.93    40.
    59.96    50.
    75.93    50.
    77.54    50.
    53.40    50.
    74.75    50.
    74.15    50.
    65.59    50.
    58.73    50.
    62.55    50.
    63.30    50.
    78.70    60.
    77.97   60.
    79.02   60.
    79.51   60.
    81.95   60.
    90.25   60.
    77.24   60.
    73.85   60.
    79.75   60.
    78.00   60.
    86.80   70.
    77.50   70.
    78.95   70.
    84.13   70.
    91.24   70.
    95.95   70.
    95.62   70.
    89.93   70.
    91.47   70.
    95.85   70.
   104.56   80.
   105.35   80.
   101.94   80.
   104.96   80.
   103.10   80.
   110.26   80.
   101.77   80.
   101.83   80.
   100.47   80.
    91.80   80.
   117.65   90.
   112.62   90.
   115.45   90.
   106.53   90.
   119.42   90.
   114.67   90.
   125.96   90.
   122.10   90.
   119.82   90.
   119.41   90.
   131.95  100.
   126.25  100.
   132.22  100.
   111.13  100.
   129.10  100.
   125.25  100.
   120.95  100.
   117.01  100.
   128.58  100.
   113.11  100.
PROC PRINT;
PROC PLOT; PLOT Y*X;
PROC GLM; MODEL Y=X;
```

```
RUNNING LINEAR REGRESSION USING PROC GLM ON SAS        10:03 SATURDAY, APRIL 27, 1985    1
                OBS         Y        X

                  1      19.41      10
                  2      21.15      10
                  3      33.62      10
                  4      25.72      10
                  5      24.60      10
                  6      23.50      10
                  7      25.59      10
                  8      29.33      10
                  9      19.37      10
                 10      20.31      10
                 11      42.73      20
                 12      40.16      20
                 13      35.11      20
                 14      41.14      20
                 15      33.52      20
                 16      31.93      20
                 17      37.53      20
                 18      32.69      20
                 19      37.65      20
                 20      31.34      20
                 21      44.58      30
                 22      52.15      30
                 23      47.24      30
                 24      46.05      30
                 25      43.71      30
                 26      43.39      30
                 27      52.15      30
                 28      45.46      30
                 29      44.42      30
                 30      37.19      30
                 31      50.57      40
                 32      61.03      40
                 33      52.24      40
                 34      55.48      40
                 35      57.02      40
                 36      49.93      40
                 37      53.09      40
                 38      53.92      40
                 39      67.53      40
                 40      61.93      40
                 41      69.95      50
                 42      75.93      50
                 43      77.54      50
                 44      63.40      50
                 45      74.75      50
                 46      74.16      50
                 47      65.59      50
                 48      63.73      50
                 49      62.55      50
                 50      63.30      50
                 51      73.70      60
                 52      77.97      60
                 53      79.02      60
                 54      78.51      60
                 55      81.85      60
                 56      90.25      60
```

```
RUNNING LINEAR REGRESSION USING PROC GLM ON SAS        10:03 SATURDAY, APRIL 27, 1985    2
            OBS          Y        X

             57        77.24     60
             58        73.35     60
             59        79.75     60
             60        73.00     60
             61        95.39     70
             62        97.59     70
             63        93.95     70
             64        94.13     70
             65        91.24     70
             66        95.95     70
             67        95.62     70
             68        97.79     70
             69        91.47     70
             70        95.35     70
             71       104.55     80
             72       105.35     80
             73       101.84     80
             74       104.95     80
             75       103.10     80
             76       110.26     80
             77       101.77     80
             78       101.38     80
             79       100.47     80
             80        91.30     80
             81       117.65     90
             82       112.62     90
             83       115.45     90
             84       106.53     90
             85       117.42     90
             86       114.67     90
             87       125.96     90
             88       122.10     90
             89       110.82     90
             90       119.41     90
             91       131.95    100
             92       126.29    100
             93       132.22    100
             94       111.13    100
             95       127.10    100
             96       125.25    100
             97       120.95    100
             98       119.01    100
             99       123.53    100
            100       112.11    100
```

```
RUNNING LINEAR REGRESSION USING PROC GLM ON SAS        10:03 SATURDAY, APRIL 27, 1985    3
PLOT OF Y*X    LEGEND: A = 1 OBS, B = 2 OBS, ETC.
```



```
                              RUNNING LINEAR REGRESSION USING PROC GLM ON SAS        10:03 SATURDAY, APRIL 27, 1935    4
                                  GENERAL LINEAR MODELS PROCEDURE

DEPENDENT VARIABLE: Y

SOURCE                  DF      SUM OF SQUARES         MEAN SQUARE        F VALUE          PR > F       R-SQUARE           C.V.
MODEL                    1     103231.67391212     103231.67391212        3395.96          0.0001       0.975469         6.3598
ERROR                   78       2596.05328789         25.49013967                       ROOT MSE                         Y MEAN
CORRECTED TOTAL         79     105827.73220000                                         5.14697669                    75.04000000

SOURCE                  DF         TYPE I SS       F VALUE      PR > F          DF         TYPE III SS       F VALUE       PR > F
X                        1     103231.67391212     3996.96      0.0001           1     103231.67391212     3996.96       0.0001

                                   T FOR H0:      PR > |T|            STD ERROR OF
PARAMETER            ESTIMATE     PARAMETER=0                         ESTIMATE

INTERCEPT         13.51633333           12.16      0.0001            1.11135244
X                  1.11351212           62.43      0.0001            0.01791912
```

4. NONLINEAR RELATIONSHIPS: NONLINEAR MODELS

4.1 Some common bivariate relationships

In words	Differential form	Integrated form	Linear form	Examples
1. The rate of change of y with respect to x is a constant.	$\frac{dy}{dx} = b$	$y = a + bx$	same	-very few in biology
2. The percentage rate of change of y with respect to x is a constant. or The rate of change of y with respect to x is proportional to y.	$\frac{dy}{y}\frac{1}{dx} = b$ $\frac{dy}{dx} = by$	$y = y_o e^{bx}$ ("exponential")	$\log y = a + bx$ ("logarithmic")	-Light intensity y at depth x in a homogeneus lake. -The amount of a compound y_1 over time x, if $y_1 \rightarrow y_2$ at a constant rate. -Population size over time x in an unlimited environment.
3. The percentage rate of change of y is proportional to the percentage rate of change of x.	$\frac{dy}{y} = b\frac{dx}{x}$	$y = Ax^b$ ("power law" or "allometric")	$\log y = a\ b(\log x)$ pp	-Body weight y related to length x for a growing animal (if shape & density do not change then b=3).
4. The rate of change of y with respect to x is proportional to the amount by which y is less than K.	$\frac{dy}{dx} = b(K-y)$	$y = K(1-e^{-bx})$ ("Von Bertalanffy" or "monomolecular")	$\log(\frac{K-y}{K}) = a - bx$	-Growth in size y with age x for many animals (a=0 if y=0 at x=0). -The amount of a compound y_2 over time x, if $y_1 \rightarrow y_2$ at a constant rate.
5. The percentage rate of change of y with respect to x is proportional to the percent of K not filled by y.	$\frac{dy}{y}\frac{1}{dx} = b\frac{(K-y)}{K}$	$y = \frac{K}{1+e^{-b(x-x_o)}}$ ("logistic")	$\log\frac{(K-y)}{y} = a - bx$	-Population size y over time x in a limited environment.
6. The percentage rate of change in the amount by which y exceeds a, with respect to x, is proportional to the amount by which y exceeds a.	$\frac{dy}{y-a}\frac{1}{dx} = b(y-a)$	$y = a - \frac{1}{bx}$	$y = a - \frac{1}{b}(x^{-1})$	-If a rate x (of mortality, say) is proportional to a stimulus y (a toxicant dose, say), then $1/x = x^{-1}$ = expected time until the event occurs (time-to-death, say).

4.2 Some common bivariate relationships: assignment/tutorial

Refer to section 4.1. We will take each of the six models in turn.

4.2.1 $y = a + bx$:

This is the model assumed by classical regression analysis, but it rarely describes relationships between variables in biology.

4.2.2 $y = y_o e^{bx}$:

Exponential growth or exponential decline of y for each unit increase in x.

Worked example: A human population grows as follows:

N	:	1000	1035	1066	1109	1147
t (yr)	:	0	1	2	3	4

The simplest sensible null hypothesis is that the population is growing at a constant % rate, which is described by the model $N = N_o e^{bt}$ or in linear form $\log N = a + bt$ where $a = \log N_o$. If we calculate the regression of log N on t, we find that $\hat{b} = 0.0343$ with .95 cl of 0.0318 to 0.0369, and $\hat{a} = 6.91$. The r^2 value, which measures the fraction of the variation in log N that is related to time t, is 0.998. Thus our model is $\log N = 6.91 + 0.0343t$, or $N = 1002e^{.034t}$. Over one year the population increases by a factor $e^{.0343}$ which is 100(1.0349 - 1) = 3.49%/yr. The lower .95 cl. on $\hat{b}$ is 0.0318, which as a % yr value is 3.23. The upper .95 cl on $\hat{b}$ is 0.0369, which as a % yr value is 3.76. Thus the .95 cl on the % rate of increase of the population are 3.23 to 3.76%/yr.

Assigned problem: First run the above worked example in MINITAB, verify that you get the same results, and then PLOT N versus t as well as log N versus t. Verify that log N versus t is an apparently linear scatter. (Use LOGE to transform to logs, and use EXPO to back-transform. See p. 47 of MINITAB manual.) Now do the following problem using MINITAB:

In 1938, Hatton scraped a rock clean of barnacles just before the annual larval set at St. Malo, France, and then at 6-month intervals he counted the number of barnacles left on the rock. The data follow (with t = 0 at time of set):

N(no./cm^2) :	15.0	8.4	4.8	1.8
t (month) :	0	6	12	18

You should be able to :

- generate a plot of N versus t and of log N versus t
- determine the regression model, both as
  log N = a + bt and as $N = N_o e^{bt}$
- put .95 cl on both $\hat{b}$ and on the %/mo, mortality
- convert the estimate of %/mo. mortality to %/yr. mortality (<u>not</u> by just multiplying by 12!)

4.2.3 $y = Ax^b$: A power law relationship between y and x.

<u>Worked example</u>: A not-so-quick biologist, not knowing the relationship between the diameter of a circle and the area of a circle, decide to determine it empirically. So he used a compass to draw many circles of various sizes, and then he measured their diameters roughly with a ruler, and he measured their areas by laying them over graph paper to count little squares. (I told you he wasn't too bright.) The data follow:

Ar(cm^2):	.5	3.4	103	28	253	.07	60	10	158	66	85	144
D (cm) :	.8	2.1	11	5.5	18	.3	8.7	3.6	14	9.2	10.4	13.6

Our biologist can at least figure out that the area of something ought to be proportional to the square of a linear measurement on the same thing, so Area = AD^b where b ought to be equal to 2. Then, log Area = a + b log D, where a = log A. If we calculate the regression of log Area on log D, we estimate $\hat{b}$ = 2.008, with to -0.1783. So our model is log Area = -0.2377 + 2.008 log D, or

Area = .788 $D^{2.01}$. An exponent of b=2 is certainly within our 0.95cl's, and the .95 cl's on A of 0.743 to 0.837 include $\pi/(2)^2 =$ 0.785. The r^2 is 0.99957.

Assigned problem: First run the above worked example in MINITAB, verify the results, and PLOT Area versus D as well as log Area versus log D. Now do the following problem using MINITAB:

Specimens of the unionid clam Anodonta grandis were collected from the Winnipeg River, in Canada, and length and volume were measured for each. The data follow:

v (ml):	11	12	18	24	27	30	36	40	43	47	54	61	76	73
L (mm):	48	53	60	62	67	70	73	74	77	79	83	86	93	94

You should be able to:

- plot V versus L and log V versus log L
- hypothesize what b should be, in a model of the form $V = AL^b$, assuming that shape does not change with growth in size.
- put .95 cl on $\hat{b}$, on a = log A, and on A.
- say whether your hypothesized value of b appearsto be correct.
- say in words what is the meaning of A in this model.
- do the following : Do the regression as 'REGR Ci ON 1 PRED. IN Cj, ST. RESIDS. IN Ck, PRED. Y IN Cm'. (See p. 66 of MINITAB manual.) Now do "PLOT Ck VERSUS Cm" to produce a plot of residuals ($\hat{y} - y_{obs}$) versus predicted values ($\hat{y}$). If the model is adequate, these should be patternless. Now do 'MPLOT Ci VERSUS Cj AND Cm VERSUS Cj" to produce a plot of the data (as log V versus log L) with the predicted values from the model log V = a + b log L also shown on the plot.

4.2.4 $y = K(1 - e^{-bx})$: Growth to an asymptote with no inflection.

Worked example: A compound A is being converted to compound B at a constant % rate, and we know that there will be 100g when it is all converted. The data collected throughout the conversion are:

B(g):	0	51	75	87	94	97
t(hr.):	0	1	2	3	4	5

In linear form the model is $\log \frac{(K - B)}{K} = a - bt$, with $K = 100$, or $\log (1 - 0.01B) = a - bt$, where $a = 0$ if $B = 0$ at $t = 0$. If we calculate the regression of $\log(1 - 0.01B)$ on t, we estimate $\hat{b} = -0.6996$ and $\hat{a} = 0.00576$, with .95 cl of -0.72 to -0.68 and - 0.057 to 0.068 respectively. The .95 cl on a include $\hat{a} = 0$. Notice that K is within a log term in the linear model which prevents us from solving for $\hat{K}$ directly if it is unknown. You could find $\hat{K}$ by trial and error - just search for the value of K that gives a minimum residual errors from the regression of log (K-B)/K on t. Alternatively you can solve for K directly by using the Walford Plot technique, to be described later.

Assigned problem: Run the above worked example in MINITAB, verify the results, and plot B versus t as well as $\log(1 - 0.01B)$ versus t. A problem based on a situation where we do not know K will be worked later, in relation to the Walford Plot technique.

4.2.5 $y = K/(1 + e^{-b(x-x_0)})$: Growth to an asymptote with an inflection halfway up.

Worked example: A dose-mortality experiment yields the following results, where M is % dead at time t :

M (%):	0	0	6	18	55	78	95	97	100
t (hr):	0	1	2	3	4	5	6	7	8

In linear form the model is $\log \frac{(K - M)}{M} = a - bt$, with $K = 100$. Values of $M = 0$ or $N = 100$ can not be used. The parameter b represents the rate of ascent, the parameter a "positions" the curve on the t-axis (allowing calculation of the $t_{m:50}$ or LT_{50}

estimate). If we caculate the regression of log (100 - M)/M on t, we estimate $\hat{b}$ = -1.302 and $\hat{a}$ = 5.158. Again K is within a log term which prevents us from solving for $\hat{K}$ directly if it is unknown, and again it would have to be found by trial and error or by the Walford Plot technique. Here confidence imits on $\hat{a}$ or $\hat{b}$ do us little good. From .95 cl on a we could only calculate 0.95 cl on M at t = 0. The best thing to do is to replicate the experiments, estimate LT_{50} for each one, and use those as replicate LT_{50} estimates for statistical tests. For this set of data the LT_{50} estimate is found by solving log(100 - 50)/50 = $\hat{a}$ + $\hat{b}$t, and it is $\hat{t}_{m=50}$ = 4.04 hr. This is $\hat{X}_o$= $\hat{t}_o$in the integrated model.

<u>Assigned problem</u>: Run the above worked example in MINITAB, verify the results, and plot M versus t as well as log(100 - M/M) versus t. Also tray the following plot, assuming that t values are in Ch M values are in Ci, and the values 0 to 8 in increments of 0.5 are in Cj. Enter the command "LET Ck = 100/(1+EXPO($\hat{b}$ * (Cj - LT_{50}))" and then "MPLOT Ci VERSUS Ch AND Ck VERSUS Cj". You will now have observed M and t values, and the fitted curve, on the same plot. A problem based on a situation where we do not know K will be worked later, using the Walford plot technique.

4.2.6 y = a - 1/bx: A hyperbolic relationship in which y is asymptotic to x = 0 and X is asymptotic to y = a.

<u>Worked example</u>: In a dose-mortality experiment the % dead at different doses is observed at each of a series of times, but even under zero dose (control) conditions the animals can not be held longer than 48 hours without mortality. We would like to estimate the dose which would cause 50% mortality (the LD_{50}) over a very long time, as would be the case with chronic exposure in the natural environment. The data from the 48-hour experiment follow:

LD (ppm):	16	13	13	12	11
t (hr) :	3	6	12	24	48

Let our model be LD = a + b'/t, where b' = −1/b. Since LD$_{50}$ becomes equal to a as t becomes very large, therefore the parameter a provides our estimate of LD$_{50}$ for a very long exposure time. As t approaches zero the LD$_{50}$ becomes very large, which implies that the organisms can withstand a very high dose for a very short time.

If we calculate the regression of LD$_{50}$ on 1/t we estimate $\hat{b}' = 14.2$ and $\hat{a} = 11.17$. The .95 cl on $\hat{a}$ are 9.93 to 12.40, which are also the .95 cl on LD$_{50}$ for a very long exposure time.

Assigned problem: Run the above worked example in MINITAB, verify the results, and plot LD$_{50}$ versus t as well as LD$_{50}$ versus 1/t.

4.2.7 Estimation of an LD$_{50}$:

There are two standard models. One is the logistic, which you used to estimate an LT$_{50}$. The only difference here would be that you would calculate the regression of log((100-M)/M) on D, the dose, at each time t. (Previously we regressed log((100 - M)/M) on t, for each does D, to obtain an LT$_{50}$). The other model is the probit or cumulative normal model. It is probably the more commonly used, but it has the disadvantage that an equation cannot be given for this model! What is needed is the integral of the normal distribution, which has no exact integral. Therefore computer programs for probit analysis do a numerical integration of a normal distribution. SAS has a probit analysis procedure of this kind.

Assigned problem: Analyse the following data using the SAS probit procedure (PROC PROBIT):

Dose (ppm):	2	3	4	5	6	7
# animals dead:	6	18	55	78	95	97
# animals exposed:	100	100	100	100	100	100

The SAS job file should be as follows:

```
TITLE  ---------------------------;
DATA   ---------;
INPUT DOSE N RES;
CARDS;
    2 100 6
    3 100 18
    4 100 55
    5 100 78
    6 100 95
    7 100 97

PROC PROBIT;
VAR DOSE N RES;
```

Those who have good memories will realize that these data are the same as for the assigned problem with the logistic, except that the variable "time" in hours, is now called "dose", in ppm. So you can compare your LD_{50} estimate in this probit analysis with the LT_{50} obtained in the logistic model analysis.

4.2.8 <u>Walford plots</u>: The growth model $y = K\ (1 - e^{-bx})$ was exemplified in section 4.2.4 by the following data set:

y (mm):	0	51	75	87	94	97
x (yr):	0	1	2	3	4	5

If we are told that K = 100, then we find b = 0.6996 by fitting the linear model $\log\ ((K - y)/K) = a - bt$, where we expect a to be zero if $y = 0$ at $x = 0$. Let us say that we do not know K <u>a priori</u>, and that we rewrite the data in the form:

y_x :	0	51	75	87	94
y_{x+1}:	51	75	87	94	97

A plot of y_{x+1} versus y_x will form a moreorless straight line $y_{x+1} = a' + b'y_x$. The parameter a' is the estimated growth during the first time unit (a year in this case), and the parameter b' is the fraction of the total growth (to the asymptote) which reamins after the first year. For these data we find $\hat{a}'=50.6$ and $\hat{b}'=0.491$. The parameters of the $y = K(1-e^{-bx})$ model are relatd to a' and b', as $b = \log b'$ and $K = a'/(1 - b')$. For these data $b = -0.71$ and $K = 99.5$, which are cose to the previous values. Confidence limits can be placed on a' and b' in the same way as reviously. Confidence lmits on b would be easy to calculate, but those on K would not be because a' and b' will not be independent of each other.

Such "y_{x+1} versus y_x" data arise very frequently. In fact we often do not have observations on y at known x values. For example, annual rings in trees, fish scales, clam shells, etc. - can be analyzed in this way. If we measure the length at an annual ring and let that be a y_x value, then we can measure the length at the next annual ring "outward" and let that be the y_{x+1} value, and so on. We need not know what the vaue of x is, and yet we can derive the growth curve for this plant or animal! Another source of such data is markrecapture studies, where we catch animals one year, measure their sizes (weight, length, or any other size measure), give them individual marks, and release them. One year later you recapture at least some of them, re-measure them, and proceed as above.

What must you assume? First of all, the time interval must be exactly the same (usually one year) for all animals. Second, if you are using annual rings you must be sure that they really <u>are</u> annual rings. Third, you must be fitting the correct Walford Plot model. If the y_{x+1} versus y_x plot is not a straight line, then $y = K(1-e^{-bx})$ is not the appropriate growth model.

model. A logistic model, $y\ K/(1 + e^{-b(x - x_o)})$, will be apropriate if the Walford plot of $(y_{x+1})^{-1}$ versus $(y_x)^{-1}$ is linear. A third growth model, commonly used for fish, is the Gompertz model. If this is appropriate, then a plot of $\log y_{x+1}$ versus $\log y_x$ will be linear. In fact there is a whole family of growth models which includes these three, and all have corresponding Walford Plot models.

<u>Assigned problem</u>: Intensive trapping of the Singapore Sling Sloth (SSS for short) was done on the N.U.S. campus. All SSS were weighed, individually marked, and released. A year later, 12 SSS were recaptured and reweighed, yielding the following data:

Animal:	1	2	3	4	5	6	7	8	9	10	11	12
1983 wt.(g):	14	17	25	30	32	40	46	50	52	58	58	70
1984 wt.(g):	53	60	49	57	67	59	67	78	72	73	78	86

Regress 1984 weight on 1983 weight. Plot the data. Estimate a' and b', and from them calculate K and b in the Von Bertalanffy model. Plot the curve $wt. = K(1 - e^{-bx})$ for age x from 0 to 10 using MINITAB commands.

4.2.9 <u>Ratio variables</u>

4.2.9.1 <u>Introduction</u>

Variables derived as the ratio of two observed variables can cause serious problems in statistical analysis. There is no problem when the denominator is a constant, as in a dose-mortality experiment where the variable "% dead" is used and is calculated as the number which have died divided by the total number at the beginning, times 100. That would amount to a change of scale, as would be the case if no. of organisms per m^2 were recorded as no. per cm^2 by dividing by 10^4. However where the denominator variable has substantial variance, estimates of the true mean of the ratio are biased and any possible correlation - between the ratio and the variables which go into the ratio - is obscured.

4.2.9.2 Worked example: A student collects a large number of hermit crabs covering a range of sizes, and expels them from their shells by applying heat to the top of the shell. Each "naked" crab is weighed and then given a choice of a range of shell sizes of the same kind (the same gastropod mollusc species) of shell. The question is, "What is the relationship between the weight of the crab and the weight of the shell that it chooses to inhabit and to carry around?"

If the student just derives the variable "ratio of shell weight to crab weight", calculates it for each combination of crab and chosen shell, and then finds the mean, standard error and .95 cl on the mean, there are two problems. The first is that a ratio variable is involved, one where the denominator is a response variable with its own substantial variance. The second is that the variance in the denominator, crab weight, may be correlated with the derived variable, ratio of shell weight to crab weight. That is, small crabs may able to carry shells that are larger in proportion to their size, compared with large crabs.

The first problem is one of a possibly biased estimate of mean ratio of shell weight to crab weight, and an inflated estimate of precision (variance, standard error and confidence limits). The second problem relates to a well-known principle in biology and in architecture: objects which must stand up above a surface must maintain their weight-to-basal area ratio as they increase in weight. Obviously an object suspended in liquid, such as a whale or a ship, is spared this problem. However for an animal on land, or for a building supported by columns, the load-bearing cross sectional area in contact with the ground increases as the square of a linear dimension (e.g. length or height) whereas the weight to be supported increases as the cube of that same linear dimension. Therefore you cannot design an elephant by describing an orders-of-magnitude larger mouse. The same is true of the architect who must design a larger version of an existing building.

Assume that the student's data are as follows:

W_s (g):	1.19	2.42	3.49	2.14	1.65	0.89	3.29	4.26	1.38	0.93
W_c (g):	1.76	2.59	6.98	3.73	1.99	1.28	4.11	10.11	2.24	2.39
(con't)	4.06	3.34	1.16	3.45	1.67	0.83	2.25	1.31	4.12	1.51
	7.04	5.08	2.29	8.79	2.15	1.72	5.75	1.52	3.87	1.42

If shell weight does increase at the same % rate as crab weight then the ratio of shell weight to crab weight will stay the same. For example, if a 4 g crab carries a 6 g shell then doubling the weight of both (a 100 % increase) would result in a 8 g crab carrying a 12 g shell. The ratio of 1.5 remains the same. Therefore we have the process "the % rate of change of one variable is proportional to the % rate of change of another variable", which leads to the power law model, and to the log-log regression model. In this case,

$$\frac{d\,W_s}{W_s} = b\,\frac{d\,W_c}{W_c}$$

$$W_s = AW_c^{\,b}$$

If the W_s/W_c ratio remains the same, then the % rate of change of W_s is equal to the % rate of change of W_c, and b should be equal to 1. If, on the other hand, our "load-bearing cross sectional area" model is applicable then b should be equal to 2/3.

In linear form our model is

$$\log W_s = a + b \log W_c ,$$

where a = log A. If we regress log W_s on log W_c, we estimate $\hat{b} = 0.742$, with .95 cl of 0.524 to 0.959, and $\hat{a} = -0.170$, with 0.95 cl of -0.453 to 0.114. The estimate of A is $\hat{A} = 0.844$, with 0.95 cl of 0.636 to 1.12.

Therefore we would conclude that b is significantly different from 1 but not from 2/3, so that the W_s/W_c ratio <u>does</u> change (decreases) as W_c increases. It changes in a manner that is compatible with the "load-bearing cross sectional area" model. We would conclude that a is not significantly different from 0, and that A is not significantly different from 1. Note that $\hat{A}$ is the estimate of the ratio of W_s to W_c at a crab weight of 1 g, and in a case where b <u>was</u> equal to 1 it would be an estimate of the ratio W_s/W_c at all values of W_c.

Now let us reformulate the model in terms of the ratio W_s/W_c.

If $$W_s = 0.844\ W_c^{0.742}$$, then

$$\frac{W_s}{W_c} = 0.844\ W_c^{-0.258}$$

A plot W_s/W_c versus W_c is attached.

Was this analysis valid? The truth is that I simulated these data under the model:

$$\frac{W_s}{W_c} = W_c^{-1/3}$$, which corresponds to

$$W_s = W_c^{2/3} .$$

4.2.10. Job Listings and Outputs.

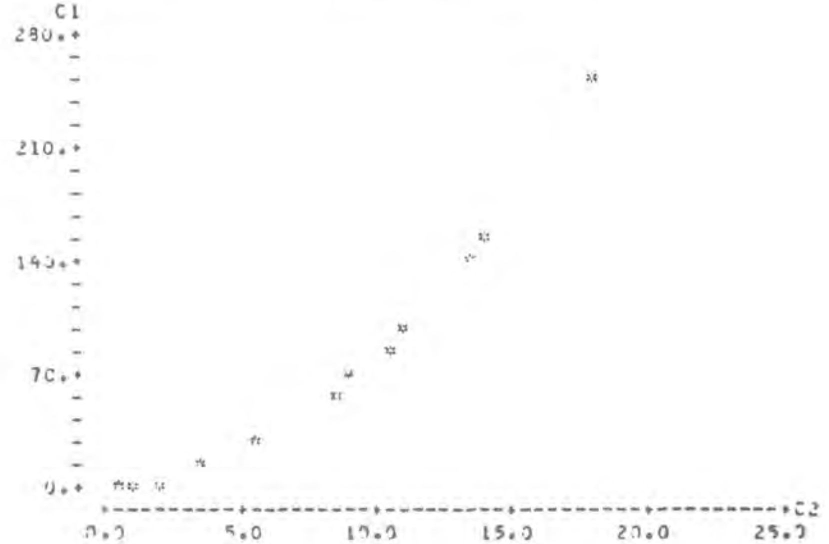
```
FILE: LA33A     MINITAB  A1  VM/SP - CONVERSATIONAL MONITOR SYSTEM                    PAGE 001

READ C1-C2
15.0  0
 8.4  6
 4.4 12
 1.3 18
LOGE C1, PUT IN C3
NOTE C3 IS THE LOGE TRANSFORMATION OF N
PRINT C1-C3
PLOT C1 VS C2
PLOT C3 VS C2
REGR C3 1 C2
EXPO 2.7352, PUT IN K1
EXPO -0.1153, PUT IN K2
LET K3=100*(K2-1)
NOTE K3 IS THE MORTALITY RATE IN % PER MONTH
LET K4=-0.1153-0.0507
NOTE K4 IS THE LOWER 95% CONFIDENCE LIMIT FOR B-HAT
LET K5=-0.1153+0.0507
NOTE K5 IS THE UPPER 95% CONFIDENCE LIMIT FOR B-HAT
PRINT K3-K5
EXPO K4, PUT IN K6
EXPO K5, PUT IN K7
LET K8=100*(K6-1)
NOTE K8 IS THE LOWER 95% CONFIDENCE LIMIT FOR MORTALITY RATE(%/MTH)
LET K9=100*(K7-1)
NOTE K9 IS THE UPPER 95% CONFIDENCE LIMIT FOR MORTALITY RATE(%/MTH)
PRINT K3-K9
LET K10=EXPO(12*(-0.1153))
LET K11=100*(K10-1)
NOTE K11 IS THE MORTALITY RATE IN % PER YEAR
PRINT K10-K11
STOP
```

```
FILE: BARNACLE OUTPUT   A1  VM/SP - CONVERSATIONAL MONITOR SYSTEM                                        PAGE 001

     NAME: KAM SUAN PHENG

1MINITAB RELEASE 81.1 *** COPYRIGHT - PENN STATE UNIV. 1981
 APRIL 29, 1985 *** NATIONAL UNIVERSITY OF SINGAPORE - LOCAL VERSION 02/12/1982
 STORAGE AVAILABLE   4800

  LINEAR REGRESSION ANALYSIS ON THE BARNACLE PROBLEM
 --
         N (#/SQ.CM)   T (MONTH)  LOGE(N)

 COLUMN        C1           C2           C3
 COUNT          4            4            4
 ROW
   1      15.0000           0.      2.70805
   2       8.4000           6.      2.12823
   3       4.8000          12.      1.56862
   4       1.8000          18.      0.58779

 --

     PLOT OF N (NO. OF BARNACLES PER SQ.CM) VERSUS T (MONTH)

           C1
        15.0+ *
            -
            -
            -
            -
        12.0+
            -
            -
            -
            -
         9.0+
            -             *
            -
            -
            -
         6.0+
            -
            -                         *
            -
            -
         3.0+
            -
            -                                     *
            -
            -
         0.0+
             +---------+---------+---------+---------+---------+C2
            0.0       5.0      10.0      15.0      20.0      25.0
```

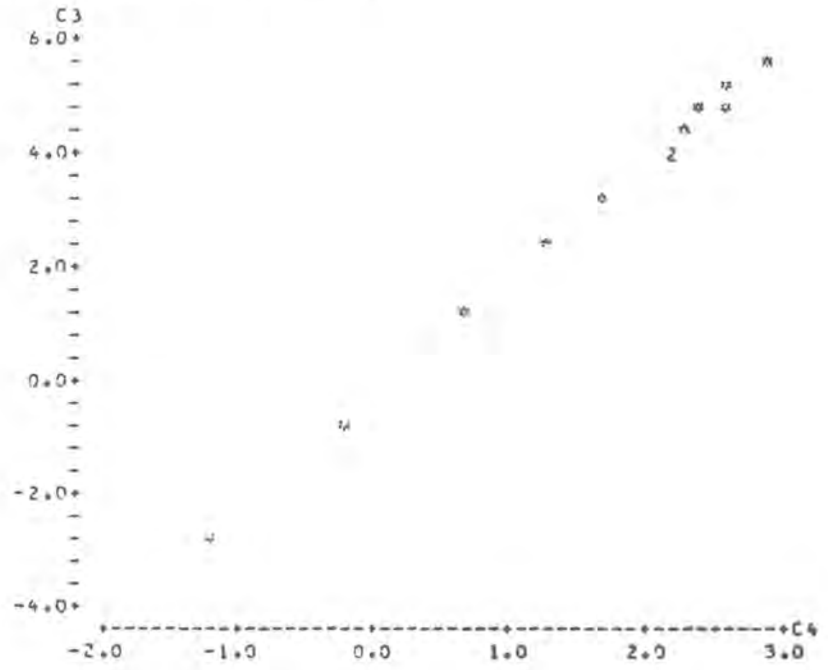
```
FILE: BARNACLE OUTPUT   A1  VM/SP - CONVERSATIONAL MONITOR SYSTEM                    PAGE 002

--
              PLOT OF LOG(N) VERSUS T

                C3
              3.00+
                  -
                  - *
                  -
                  -
              2.40+
                  -
                  -           *
                  -
                  -
              1.80+
                  -
                  -                       *
                  -
                  -
              1.20+
                  -
                  -
                  -
                  -
              0.60+                                   *
                  +---------+---------+---------+---------+---------+C2
                 0.0       5.0      10.0      15.0      20.0      25.0

   LINEAR REGRESSION ANALYSIS (WITH DATA TRANSFORMATION)
   MODEL IS:                      LOG(N) = A + BT     -------- (1)
   THE EXPONENTIAL FORM IS:          N = N EXP(BT) -------- (2)
   WHERE                         LOG(N ) = A          -------- (3)
THE REGRESSION EQUATION IS
LOG(N) = 2.79 - 0.115T

                                               ST. DEV.     T-RATIO =
       COLUMN       COEFFICIENT               OF COEF.     COEF/S.D.
       INTERCEPT        2.7862                 0.1322          21.07
       SLOPE          -0.11534                0.01178          -9.79

THE ST. DEV. OF Y ABOUT REGRESSION LINE IS
S = 0.1580
WITH (    4- 2) =    2 DEGREES OF FREEDOM

R-SQUARED = 98.0 PERCENT
R-SQUARED = 96.9 PERCENT, ADJUSTED FOR D.F.

ANALYSIS OF VARIANCE

 DUE TO        DF           SS         MS=SS/DF
REGRESSION      1      2.39459          2.39459
RESIDUAL        2      0.04994          0.02497
TOTAL           3      2.44453
```

```
 DURBIN-WATSON STATISTIC = 2.27
 --
   K1 = EXP(A) =    16.2193
 --
   FROM EQUATION (3) ABOVE,  M  = EXP(A) = K1

THEREFORE THE REGRESSION MODEL IN THE EXPONENTIAL FORM IS:
                 N = 16.22EXP(-0.1153T)

   K2 = EXP(-0.1153) = 0.8911
 --
 --
   K3 =  -10.8901
   K3 IS THE ESTIMATED MORTALITY RATE IN % PER MONTH

   K4        -0.156009
   K5       -0.0546000
   K4 AND K5 ARE THE LOWER AND UPPER 95% CONFIDENCE LIMITS FOR B-HAT

 --
   K6 = EXP(K4) =    0.7509
 --
   K7 = EXP(K5) =    1.0435
 --
 --
   K8          -15.2954
   K9          -6.23575
   K8 AND K9 ARE THE LOWER AND UPPER 95% CONFIDENCE LIMITS FOR
   THE MORTALITY RATE (%/MONTH)
 --
  TO CALCULATE THE MORTALITY RATE IN #/YEAR
  THE REGRESSION MODEL IS     LOG(N) = A + 12BT'
  WHERE      T' = 12T;  T' IS TIME IN YEAR

   K10 = EXP(12*(-0.1153)) =  0.250675
   K11          -74.9325
   K11 IS THE MORTALITY RATE = -74.9% PER YEAR
*** MINITAB *** STATISTICS DEPT * PENN STATE UNIV. * RELEASE 81.1 *
 STORAGE AVAILABLE    4000
```

```
FILE: LAB301   OUTPUT   A1  VM/SP - CONVERSATIONAL MONITOR SYSTEM                                PAGE 001

MINITAB RELEASE 81.1 *** COPYRIGHT - PENN STATE UNIV. 1981
 APRIL 30, 1985 *** NATIONAL UNIVERSITY OF SINGAPORE - LOCAL VERSION 02/12/1982
 STORAGE AVAILABLE   4900

 LINEAR REGRESSION ANALYSIS OF AREA OF CIRCLE ON DIAMETER
      MODEL:  LOG(A ) = A + BLOG(D)
 --
 COLUMN         C1          C2          C3          C4
 COUNT          12          12          12          12
 ROW            A           D        LOG(A )      LOG(D)
   1          0.500      0.8000    -0.69315    -0.22314
   2          3.400      2.1000     1.22377     0.74194
   3        103.000     11.0000     4.63473     2.39790
   4         28.000      5.5000     3.33220     1.70475
   5        253.000     18.0000     5.53339     2.89037
   6          0.070      0.3000    -2.65926    -1.20397
   7         60.000      8.7000     4.09434     2.16332
   8         10.000      3.6000     2.30258     1.28093
   9        158.000     14.0000     5.06260     2.63906
  10         66.000      9.2000     4.18965     2.21920
  11         85.000     10.4000     4.44265     2.34181
  12        144.000     13.6000     4.96981     2.61007

 --
   PLOT OF AREA OF CIRCLE (A ) VERSUS DIAMETER (D)
```



PLOT OF LOG(A) VERSUS LOG(D)



```
--

THE REGRESSION EQUATION IS
LOG(A) = - 0.238 +  2.01LOG(D)

                                        ST. DEV.     T-RATIO =
       COLUMN       COEFFICIENT         OF COEF.     COEF/S.D.
       INTERCEPT     -0.23767           0.02664        -8.92
       SLOPE          2.00324           0.01311       153.19

THE ST. DEV. OF Y ABOUT REGRESSION LINE IS
S = 0.05512
WITH (  12- 2) =  10 DEGREES OF FREEDOM

R-SQUARED =100.0 PERCENT
R-SQUARED =100.0 PERCENT, ADJUSTED FOR D.F.
```

```
FILE: LAB3B1   OUTPUT   A1  VM/SP - CONVERSATIONAL MONITOR SYSTEM                    PAGE 003


 ANALYSIS OF VARIANCE

  DUE TO      DF          SS       MS=SS/DF
 REGRESSION    1    71.29090       71.29090
 RESIDUAL     10     0.03039        0.00304
 TOTAL        11    71.32124

            X1           Y     PRED. Y    ST.DEV.
 ROW        C4          C3       VALUE    PRED. Y   RESIDUAL     ST.RES.
   4      1.70      3.3322      3.1858     0.0159     0.1464       2.77R
   6     -1.20     -2.6593     -2.6555     0.0404    -0.0037      -0.10 X

 R DENOTES AN OBS. WITH A LARGE ST. RES.
 X DENOTES AN OBS. WHOSE X VALUE GIVES IT LARGE INFLUENCE.

 DURBIN-WATSON STATISTIC = 1.70
 --

 *** MINITAB *** STATISTICS DEPT * PENN STATE UNIV. * RELEASE 81.1 *
 STORAGE AVAILABLE    4800
```

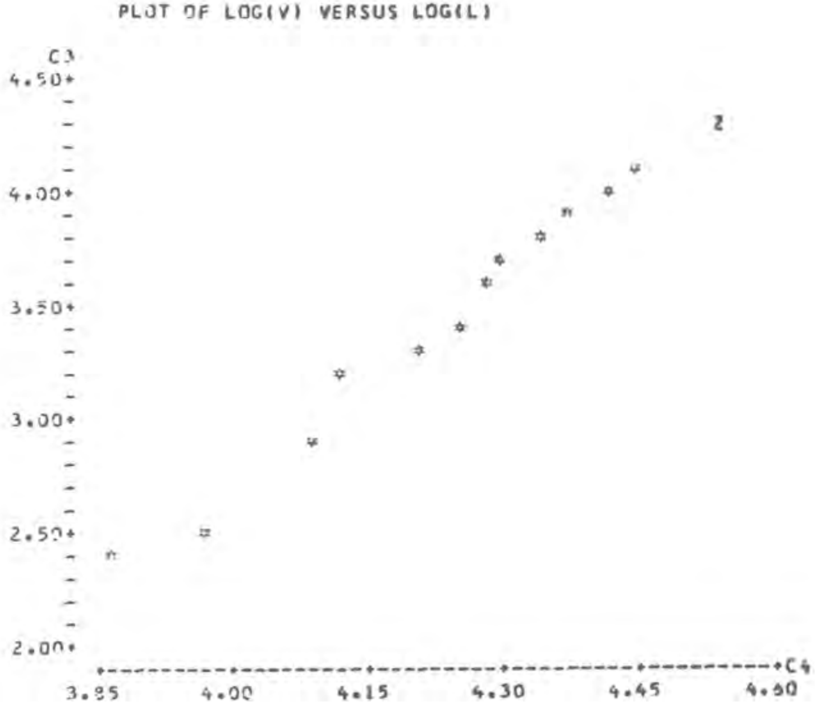
```
FILE: PLOT     OUTPUT   A1  VM/SP - CONVERSATIONAL MONITOR SYSTEM                              PAGE 001

  EXAMPLE OF PLOT PROCEDURE USING FORTRAN

 THE DATA AS READ IN ARE:
      X                    Y
   10.0000000          19.4100035
   20.0000000          42.9299925
   30.0000000          44.5300013
   40.0000000          60.5699981
   50.0000000          69.7500057
   60.0000000          73.6999969
   70.0000000          96.9000030
   80.0000000          104.5599937
   90.0000000          117.6499993
   100.000000          131.9499975


   HORIZONTAL AXIS IS DIMENSION  1
   VERTICAL AXIS IS DIMENSION    2


          HORIZONTAL AXIS

   MINIMUM VALUE=        10.00000
   MAXIMUM VALUE=       100.00000
   SCALING UNIT =         1.50000
   ONE TICK=      0.000

          VERTICAL AXIS

   MINIMUM VALUE=        19.41000
   MAXIMUM VALUE=       131.95000
   SCALING UNIT =         5.62700
   ONE TICK=      0.000

   OVERLAPPING OBJECTS (NOT PLOTTED)

   ID.NUMBER          COORDINATES
```

```
FILE: REGR      OUTPUT   A1  VM/SP - CONVERSATIONAL MONITOR SYSTEM                                    PAGE 001

  EXAMPLE OF LINEAR REGRESSION ANALYSIS (WITH TRANSFORMATION)
  USING FORTRAN
       MODEL: LOG(Y) = A + BLOG(X)


THE DATA AS READ IN, BEFORE ANY TRANSFORMATION, ARE:
     Y              Y
    0.800          0.500
    2.100          3.400
   11.000        103.000
    5.500         28.000
   10.900        253.000
    0.300          0.070
    8.700         60.000
    3.600         10.000
   14.000        158.000
    9.200         66.000
   10.400         85.000
   13.600        144.000

THE DATA AFTER TRANSFORMATION, IF ANY, ARE:
     X              Y
   -0.223         -0.693
    0.742          1.224
    2.398          4.635
    1.705          3.332
    2.390          5.533
   -1.204         -2.659
    2.163          4.094
    1.281          2.303
    2.639          5.063
    2.219          4.190
    2.342          4.443
    2.610          4.970

   X MEAN=     1.63   Y MEAN=     3.04
X VARIANCE=     1.61   Y VARIANCE=     6.48   XY COVARIANCE =     3.23


 THE REGRESSION LINE IS Y=  -0.2377 +     2.0032X


 THE ANALYSIS OF VARIANCE TABLE IS:

 SOURCE  SUM OF SQUARES  MEAN SQUARE  F-STATISTIC
    1         71.29         71.29        ********
   10          0.03          0.00
 -----     ---------
   11         71.32

 R-SQUARED=.99958   PERCENT R-SQUARED=99.96


Y-PREDICTEDS AND Y-RESIDUALS FOLLOW.
```

```
READ C1-C2
11 48
12 53
18 60
24 63
27 67
30 70
36 73
40 74
43 77
47 79
54 83
61 36
76 93
73 94
LOGE C1, PUT IN C3
NOTE C3 IS THE LOGE TRANSFORMATION OF V
LOGE C2, PUT IN C4
NOTE C4 IS THE LOGE TRANSFORMATION OF L
PRINT C1-C4
PLOT C1 VS C2
PLOT C3 VS C4
REGR C3 1 C4, STD RESIDUALS IN C5, PREDICTED LOG(V) IN C6
WIDTH 100, 50
MPLOT C3 VS C4 AND C6 VS C4
STOP
```

```
1MINITAB RELEASE 81.1 *** COPYRIGHT - PENN STATE UNIV. 1981
 MAY  2, 1985 *** NATIONAL UNIVERSITY OF SINGAPORE - LOCAL VERSION 02/12/1982
  LINEAR REGRESSION OF CLAM DATA
        MODEL: LOG(V) = A + BLOG(L)    WHERE V IS VOLUME IN ML
                                             L IS LENGTH IN MM
 --
 COLUMN        C1          C2          C3          C4
 COUNT         14          14          14          14
 ROW           V           L         LOG(V)      LOG(L)
   1          11.         48.       2.39790     3.87120
   2          12.         53.       2.48491     3.97029
   3          18.         60.       2.89037     4.09434
   4          24.         62.       3.17805     4.12713
   5          27.         67.       3.29584     4.20469
   6          30.         70.       3.40120     4.24850
   7          36.         73.       3.58352     4.29046
   8          40.         74.       3.68888     4.30406
   9          43.         77.       3.76120     4.34381
  10          47.         79.       3.85015     4.36945
  11          54.         83.       3.98898     4.41884
  12          61.         86.       4.11087     4.45435
  13          76.         93.       4.33073     4.53260
  14          73.         94.       4.29046     4.54329

 --                       PLOT OF VOLUME (V) VERSUS LENGTH (L)
        C1
      85.+
         -
         -
         -                                                 *
         -                                                  *
      70.+
         -
         -
         -                                          *
         -
      55.+                                         *
         -
         -
         -                                     *
         -                                   *
      40.+                                *
         -                               *
         -
         -                            *
         -                         *
      25.+                    *
         -
         -                  *
         -
         -         *
      10.+    *
           +---------+---------+---------+---------+---------+C2
          45.       55.       65.       75.       85.       95.
```

```
--
```

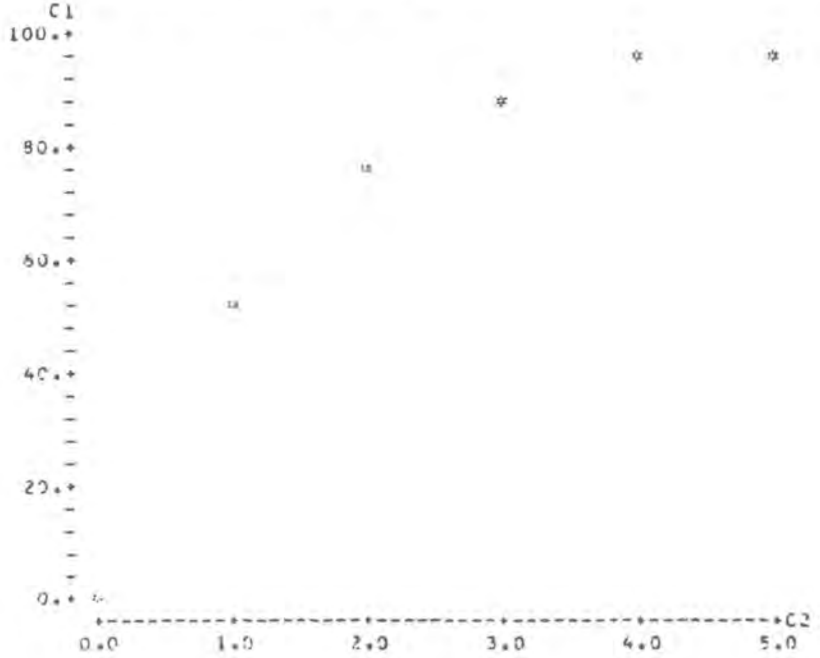


```
--

THE REGRESSION EQUATION IS
LOG(V) = -  9.53 +  3.06 LOG(L)

                                              ST. DEV.    T-RATIO =
        COLUMN      COEFFICIENT               OF COEF.    COEF/S.D.
        INTERCEPT      -9.5322                 0.4002       -23.82
        SLOPE           3.05562               0.09364        32.64

THE ST. DEV. OF Y ABOUT REGRESSION LINE IS
S = 0.06763
WITH (  14- 2) =  12 DEGREES OF FREEDOM

R-SQUARED = 98.9 PERCENT
R-SQUARED = 98.8 PERCENT, ADJUSTED FOR D.F.
```

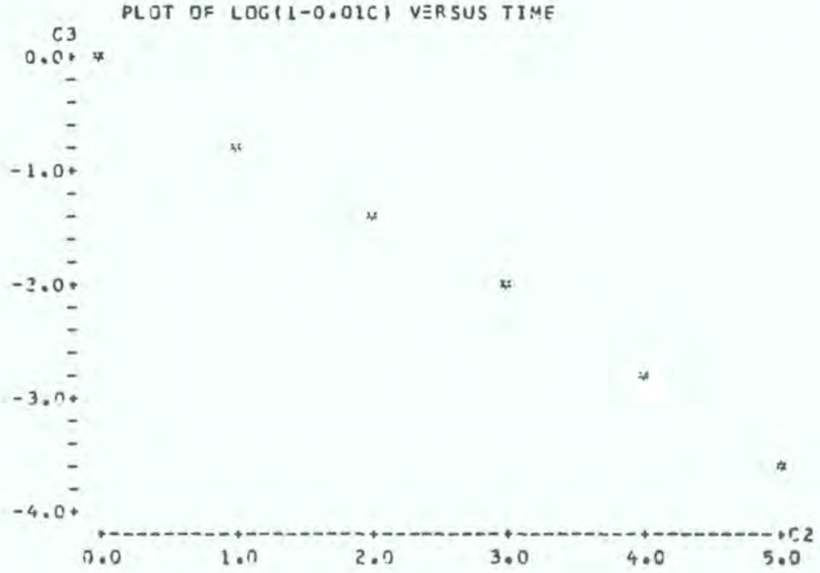
```
FILE: CLAM     OUTPUT   A1  VM/SP - CONVERSATIONAL MONITOR SYSTEM                                   PAGE 003

 ANALYSIS OF VARIANCE

  DUE TO       DF          SS         MS=SS/DF
 REGRESSION     1     4.873100        4.873100
 RESIDUAL      12     0.054880        0.004573
 TOTAL         13     4.927977

               X1           Y      PRED. Y     ST.DEV.
 ROW           C4          C3       VALUE     PRED. Y    RESIDUAL        ST.RES.
   1         3.87      2.3979      2.3006      0.0414      0.0973          1.32 X
   2         3.97      2.4849      2.6335      0.0333     -0.1186         -2.02R

 R DENOTES AN OBS. WITH A LARGE ST. RES.
 X DENOTES AN OBS. WHOSE X VALUE GIVES IT LARGE INFLUENCE.

 DURBIN-WATSON STATISTIC = 2.05
```

```
FILE: LAB3C1   OUTPUT   A1  VM/SP - CONVERSATIONAL MONITOR SYSTEM                                    PAGE 001

1MINITAB RELEASE 81.1 *** COPYRIGHT - PENN STATE UNIV. 1981
 APRIL 30, 1985 *** NATIONAL UNIVERSITY OF SINGAPORE - LOCAL VERSION 02/12/1982
 STORAGE AVAILABLE   4900

  LINEAR REGRESSION OF GENERATION OF COMPOUND C FROM COMPOUND A ON TIME
         MODEL: Y = K(1-EXP(-BC)    K = 100
            OR  LOG(1 - 0.001C) = A + BT
 --
 COLUMN        C1          C2          C3
 COUNT          6           6           6
 ROW            C           T     LOG(1-0.01C)
   1           0.          0.       0.0
   2          51.          1.     -0.71335
   3          75.          2.     -1.39629
   4          87.          3.     -2.04022
   5          94.          4.     -2.81341
   6          97.          5.     -3.50655

 --
```

PLOT OF LOG(1-0.01C) VERSUS TIME



```
 --
```

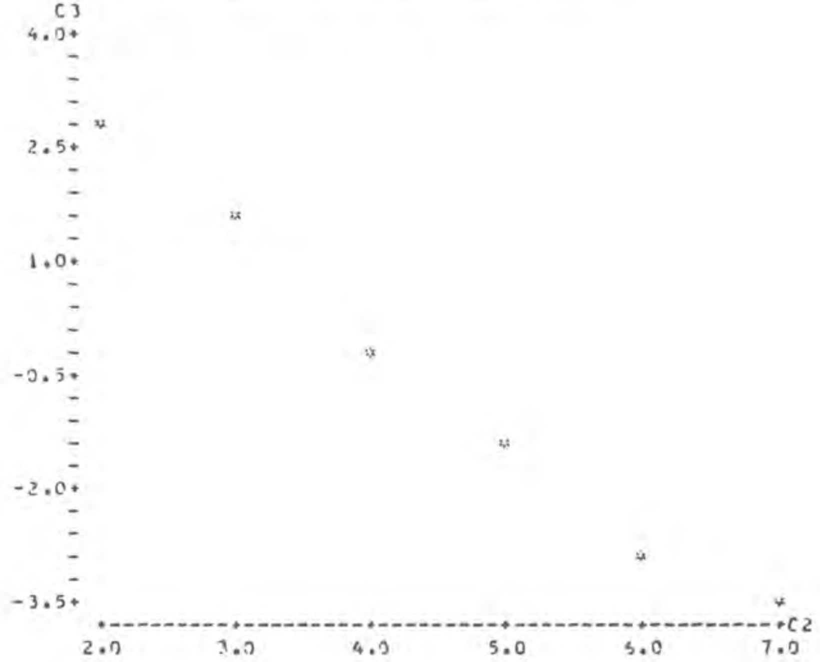


```
--

THE REGRESSION EQUATION IS
LOG(1-0.01C) =  0.0058 - 0.700 T

                                             ST. DEV.     T-RATIO =
        COLUMN      COEFFICIENT              OF COEF.     COEF/S.D.
        INTERCEPT     0.00576                0.02257          0.26
        SLOPE        -0.697524              0.007456        -93.84

THE ST. DEV. OF Y ABOUT REGRESSION LINE IS
S = 0.03119
WITH (    6- 2) =    4 DEGREES OF FREEDOM

R-SQUARED =100.0 PERCENT
R-SQUARED = 99.9 PERCENT, ADJUSTED FOR D.F.

ANALYSIS OF VARIANCE

 DUE TO       DF            SS         MS=SS/DF
REGRESSION     1      8.565739         8.565739
RESIDUAL       4      0.003891         0.000973
TOTAL          5      8.569686

DURBIN-WATSON STATISTIC = 2.17
```

```
READ C1-C2
  6 2
 18 3
 55 4
 79 5
 95 6
 97 7
LET C3=LOGE((100-C1)/C1)
NOTE C3 IS THE LOGE((100-Y)/Y); K=100
PRINT C1-C3
PLOT C1 VS C2
PLOT C3 VS C2
REGR C3 1 C2
SET C4
0 0.5 1.0 1.5 2.0 2.5 3.0 3.5 4.0 4.5 5.0 5.5 6.0 6.5 7.0 7.5 8.0
LET C5=100/(1+EXPO((-1.30244)*(C4-4.0367)))
WIDTH 100, 50
MPLOT C1 VS C2 AND C5 VS C4
STOP
```

```
FILE: DJSE     OUTPUT   A1  VM/SP - CONVERSATIONAL MONITOR SYSTEM                                   PAGE 001

IMINITAB RELEASE 81.1 *** COPYRIGHT - PENN STATE UNIV. 1981
 APRIL 30, 1985 *** NATIONAL UNIVERSITY OF SINGAPORE - LOCAL VERSION 02/12/1982
 STORAGE AVAILABLE   4900

  LINEAR REGRESSION ANALYSIS OF % MORTALITY ON TIME
       MODEL: LOG((K-M)/M) = A + BT       K = 100
 --
 COLUMN        C1           C2            C3
 COUNT          6            6             6
 ROW
   1           6.           2.       2.75154
   2          18.           3.       1.51635
   3          55.           4.      -0.20067
   4          78.           5.      -1.26567
   5          95.           6.      -2.94444
   6          97.           7.      -3.47610

 --
                PLOT OF % MORTALITY VERSUS TIME

        C1
      100.+
          -                                         *         *
          -
          -
          -
       80.+                               *
          -
          -
          -
          -
       60.+
          -                     *
          -
          -
          -
       40.+
          -
          -
          -
          -
       20.+           *
          -
          - *
          -
        0.+
            +---------+---------+---------+---------+---------+C2
           2.0       3.0       4.0       5.0       6.0       7.0


 --
```

PLOT OF LOG((100-M)/M) VERSUS TIME



```
--

THE REGRESSION EQUATION IS
LOG((K-M)/M) =     5.26 -  1.30 T
                                          ST. DEV.     T-RATIO =
        COLUMN      COEFFICIENT           OF COEF.     COEF/S.D.
        INTERCEPT      5.2579              0.3621        14.52
        SLOPE         -1.30244             0.07524      -17.31

THE ST. DEV. OF Y ABOUT REGRESSION LINE IS
S = 0.3148
WITH (   6- 2) =    4 DEGREES OF FREEDOM

R-SQUARED = 98.7 PERCENT
R-SQUARED = 98.4 PERCENT, ADJUSTED FOR D.F.

ANALYSIS OF VARIANCE

 DUE TO       DF            SS        MS=SS/DF
REGRESSION    1      29.68620      29.68620
RESIDUAL      4       0.39628       0.09907
TOTAL         5      30.08247

DURBIN-WATSON STATISTIC = 2.44
```

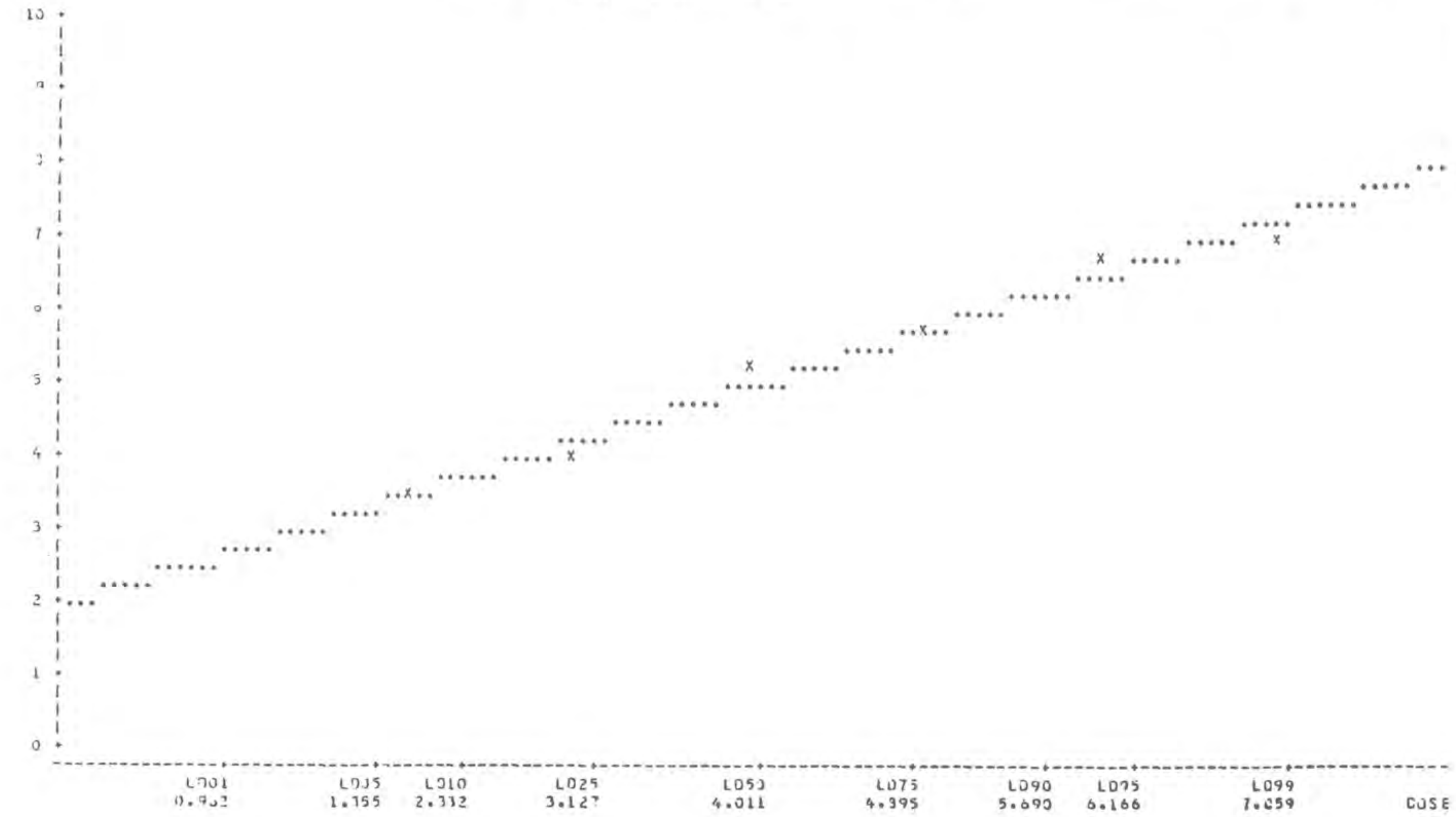
```
FILE: LAB3E     OUTPUT    A1  VM/SP - CONVERSATIONAL MONITOR SYSTEM                            PAGE 001

1MINITAB RELEASE 81.1 *** COPYRIGHT - PENN STATE UNIV. 1981
 APRIL 30, 1985 *** NATIONAL UNIVERSITY OF SINGAPORE - LOCAL VERSION 02/12/1982
 STORAGE AVAILABLE   4900

  LINEAR REGRESSION OF LD50 ON RECIPROCAL OF TIME
        MODEL:  LD50 = A + B'/T
 --
 COLUMN       C1          C2          C3
 COUNT         5           5           5
 ROW
   1         16.          3.    0.333333
   2         13.          6.    0.166667
   3         13.         12.    0.083333
   4         12.         24.    0.041667
   5         11.         48.    0.020833

 --
                    PLOT OF LD50 VERSUS TIME

                C1
             16.0+    *
                 -
                 -
                 -
                 -
             15.0+
                 -
                 -
                 -
                 -
             14.0+
                 -
                 -
                 -
                 -
             13.0+       *     *
                 -
                 -
                 -
                 -
             12.0+                         *
                 -
                 -
                 -
                 -
             11.0+                                                 *
                   +---------+---------+---------+---------+---------+C2
                   0.       10.       20.       30.       40.       50.
```

```
--
                  PLOT OF LD50 VERSUS RECIPROCAL OF TIME

          C1
        16.0+                                        *
            -
            -
            -
            -
        15.0+
            -
            -
            -
            -
        14.0+
            -
            -
            -
            -
        13.0+          *          *
            -
            -
            -
            -
        12.0+     *
            -
            -
            -
            -
        11.0+   *
             +---------+---------+---------+---------+---------+C3
           0.0     0.080     0.160     0.240     0.320     0.400

--

THE REGRESSION EQUATION IS
Y =     11.2 +  14.2 (1/T)

                                                ST. DEV.     T-RATIO =
       COLUMN       COEFFICIENT                 OF COEF.     COEF/S.D.
       INTERCEPT      11.1667                    0.3889         28.72
       SLOPE           14.194                     2.250          5.23

THE ST. DEV. OF Y ABOUT REGRESSION LINE IS
S = 0.5742
WITH (    5- 2) =    3 DEGREES OF FREEDOM

R-SQUARED = 92.9 PERCENT
R-SQUARED = 90.6 PERCENT, ADJUSTED FOR D.F.
```

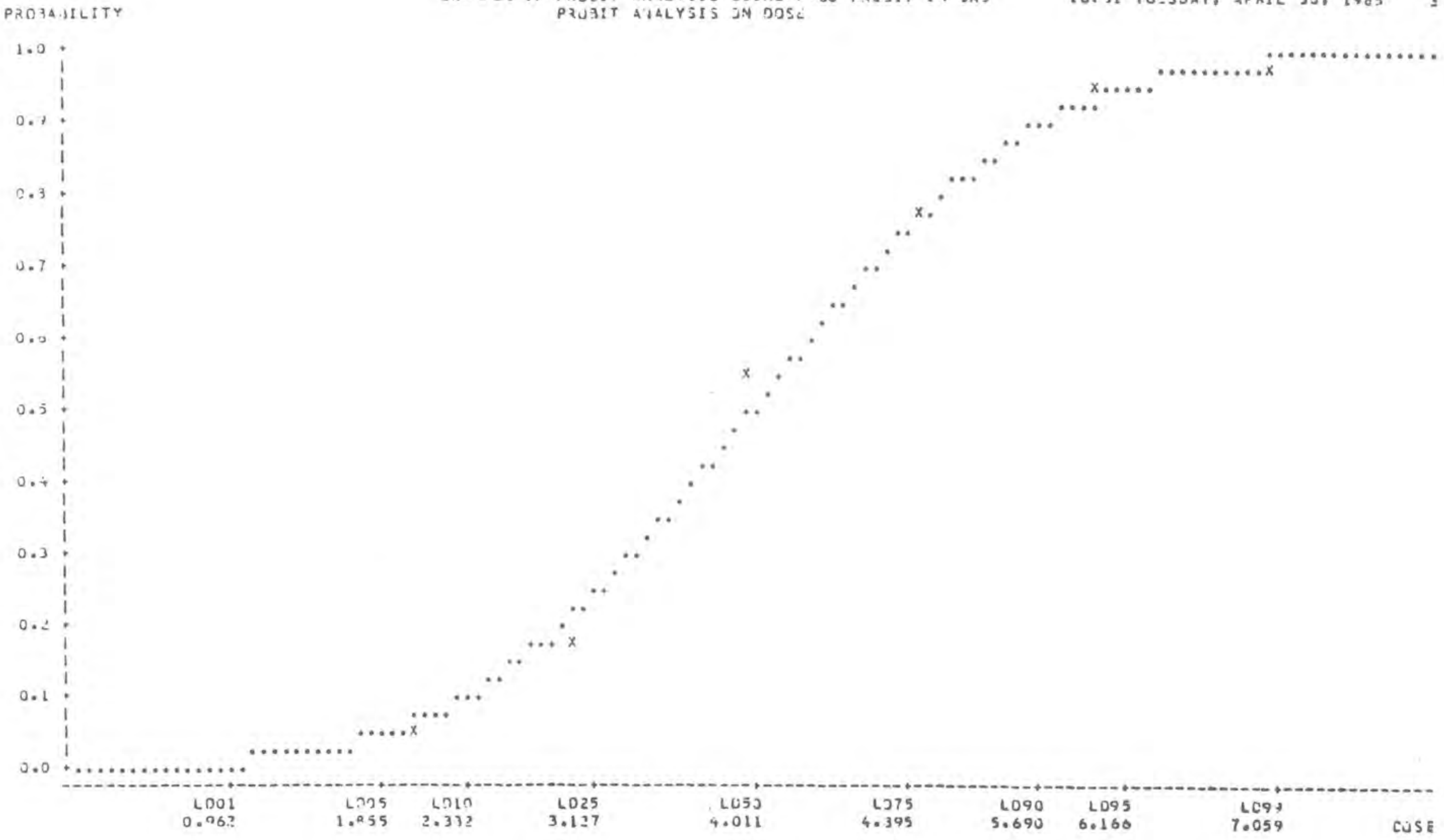
```
 ANALYSIS OF VARIANCE

  DUE TO       DF           SS        MS=SS/DF
 REGRESSION     1      13.0107         13.0107
 RESIDUAL       3       0.9892          0.3297
 TOTAL          4      14.0000

    1      0.333      16.000      15.893       0.528       0.102        0.45 X

 DURBIN-WATSON STATISTIC = 2.49
 --

 *** MINITAB *** STATISTICS DEPT * PENN STATE UNIV. * RELEASE 81.1 *
 STORAGE AVAILABLE      4800
```

```
               EXAMPLE OF PROBIT ANALYSIS USING PROC PROBIT IN SAS        16:31 TUESDAY, APRIL 30, 1985   1
                               PROBIT ANALYSIS ON DOSE

     ITERATION           INTERCEPT            SLOPE                 MU               SIGMA

         0               2.04634315         0.72371477          4.05324025        1.37227965
         1               1.94459905         0.75154969          4.01204350        1.31311195
         2               1.93915063         0.76314104          4.01035673        1.31037377
         3               1.93913995         0.76314406          4.01025433        1.31036858

COVARIANCE MATRIX                                              COVARIANCE MATRIX
                   INTERCEPT            SLOPE                                         MU                 SIGMA

INTERCEPT          0.04960331        -0.01069361               MU              0.00758349          -0.00069754
SLOPE             -0.01069361         0.00253393               SIGMA          -0.00069754           0.00753283

                  CHI-SQ =     5.6015 WITH        4 DF       PROB > CHI-SQ = 0.2309

NOTE: SINCE THE CHI-SQUARE IS SMALL (P > 0.10), FIDUCIAL LIMITS WILL BE COMPUTED USING A T VALUE OF 1.96 .
```

```
EXAMPLE OF PROBIT ANALYSIS USING PROC PROBIT IN SAS        16:31 TUESDAY, APRIL 30, 1985    2
                  PROBIT ANALYSIS ON DOSE
```



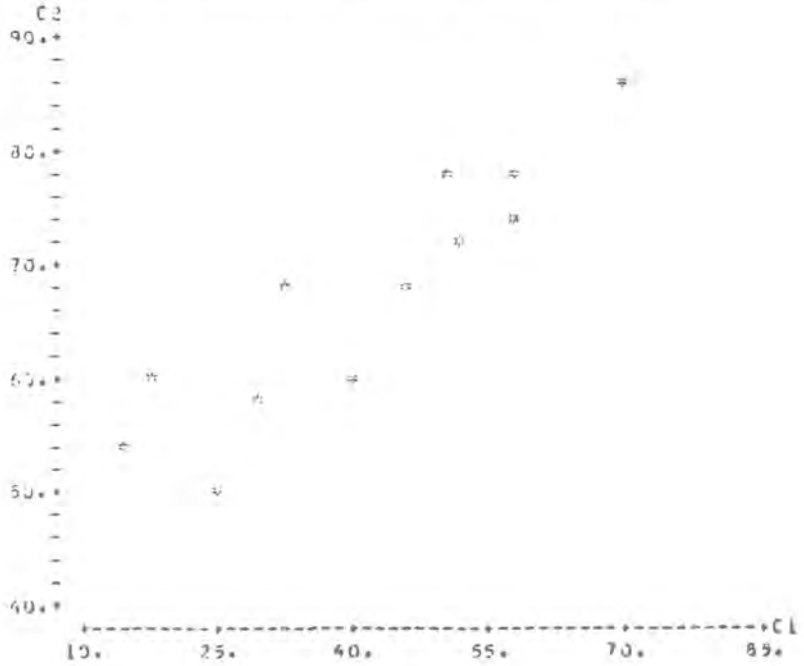
EXAMPLE OF PROBIT ANALYSIS USING PROC PROBIT IN SAS 16:31 TUESDAY, APRIL 30, 1985 3
PROBIT ANALYSIS ON DOSE



```
             EXAMPLE OF PROBIT ANALYSIS USING PROC PROBIT IN SAS        15:31 TUESDAY, APRIL 30, 1985    4
                           PROBIT ANALYSIS ON DOSE

PROBABILITY               DOSE               95 PERCENT FIDUCIAL LIMITS
                                                LOWER                 UPPER

   0.01                0.95243159            0.45225901            1.36262062
   0.02                1.31063580            0.85922092            1.68244036
   0.03                1.54532179            1.11694124            1.88590538
   0.04                1.71681032            1.31050498            2.03924393
   0.05                1.85549033            1.45772272            2.16420334
   0.06                1.97352337            1.60135134            2.27075232
   0.07                2.07702465            1.71835809            2.36433484
   0.08                2.16969321            1.82298150            2.44826844
   0.09                2.25397157            1.91300435            2.52473035
   0.10                2.33154093            2.00535476            2.59523261
   0.15                2.65274509            2.36556371            2.88357464
   0.20                2.90302032            2.64971401            3.12584613
   0.25                3.12702456            2.89145391            3.32772371
   0.30                3.32359513            3.10648032            3.51297508
   0.35                3.50594300            3.30357645            3.68660425
   0.40                3.67437675            3.48830643            3.85375074
   0.45                3.84619215            3.66459639            4.01791535
   0.50                4.01035493            3.83547106            4.19207762
   0.55                4.17551751            4.00362409            4.34897154
   0.60                4.34283271            4.17165921            4.52133199
   0.65                4.51376666            4.34245079            4.70246589
   0.70                4.69301279            4.51953268            4.89620881
   0.75                4.89468500            4.70771040            5.10820936
   0.80                5.11368835            4.91424937            5.34724756
   0.85                5.36376458            5.15197036            5.62494847
   0.90                5.69015973            5.44749356            5.98701615
   0.91                5.76773809            5.51941402            6.07394816
   0.92                5.85201645            5.59520212            6.16855521
   0.93                5.94463502            5.67964015            6.27276360
   0.94                6.04819129            5.77364098            6.39935234
   0.95                6.16521933            5.88061342            6.52255739
   0.96                6.30489834            6.00601055            6.67733802
   0.97                6.47538767            6.15980932            6.87244043
   0.98                6.70202235            6.36374303            7.12965257
   0.99                7.05922778            6.68422119            7.53399566
```

```
1MINITAB RELEASE 81.1 *** COPYRIGHT - PENN STATE UNIV. 1981
 MAY  2, 1985 *** NATIONAL UNIVERSITY OF SINGAPORE - LOCAL VERSION 02/12/1982
 STORAGE AVAILABLE   4900

  EXAMPLE OF USING WALFORD PLOT FOR OBTAINING THE
  "VON BERTALANFFY" GROWTH EQUATION FOR SSS
 --
 COLUMN        C1          C2
 COUNT         12          12
 ROW       W AT 1983   W AT 1984
   1          14.         53.
   2          17.         60.
   3          25.         49.
   4          30.         57.
   5          32.         67.
   6          40.         59.
   7          46.         67.
   8          50.         78.
   9          52.         72.
  10          58.         73.
  11          58.         78.
  12          70.         85.

 --
```

PLOT OF WEIGHT IN 1984 VERSUS WEIGHT IN 1983



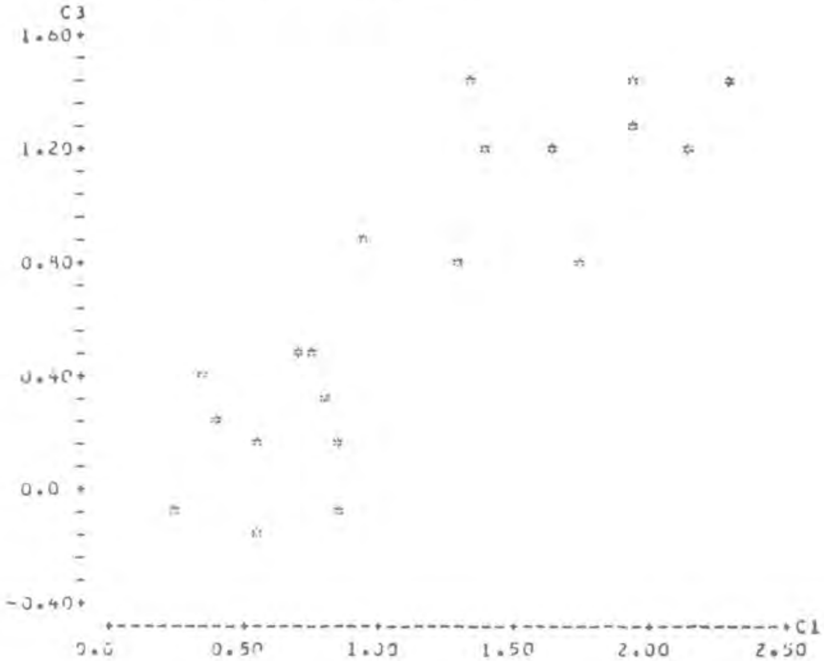
```
 --

 THE REGRESSION EQUATION IS
 Y  =    43.2 + 0.571 X

                                                ST. DEV.     T-RATIO =
        COLUMN       COEFFICIENT                OF COEF.     COEF/S.D.
        INTERCEPT        43.135                   4.049          10.67
        SLOPE            0.57059                0.09132           6.25

 THE ST. DEV. OF Y ABOUT REGRESSION LINE IS
 S = 5.333
 WITH (  12- 2) =  10 DEGREES OF FREEDOM

 R-SQUARED = 79.6 PERCENT
 R-SQUARED = 77.6 PERCENT, ADJUSTED FOR D.F.

 ANALYSIS OF VARIANCE

  DUE TO       DF             SS       MS=SS/DF
 REGRESSION    1        1110.54        1110.54
 RESIDUAL      10        284.38          28.44
 TOTAL         11       1394.92


 DURBIN-WATSON STATISTIC = 2.51
 --
 ITERATION    1   Y =      0.554054 X +    43.301804
 ITERATION    2   Y =      0.600433 X +    42.356764
 SLOPE =        0.6055
 LEVEL =       42.2450
 --
 --

 THE REGRESSION EQUATION IS
 Y =      51.9 +0.0507 X1 +0.0064 X2

                                                ST. DEV.     T-RATIO =
        COLUMN       COEFFICIENT                OF COEF.     COEF/S.D.
        --                51.976                  8.609           6.03
 X1     C1                0.0507                 0.4542           0.11
 X2     C3              0.006416               0.005629           1.14

 THE ST. DEV. OF Y ABOUT REGRESSION LINE IS
 S = 5.254
 WITH (  12- 3) =   9 DEGREES OF FREEDOM

 R-SQUARED = 82.2 PERCENT
 R-SQUARED = 78.2 PERCENT, ADJUSTED FOR D.F.
```

```
FILE: SSS      OUTPUT   A1  VM/SP - CONVERSATIONAL MONITOR SYSTEM                                  PAGE 003

 ANALYSIS OF VARIANCE

  DUE TO      DF          SS        MS=SS/DF
 REGRESSION    2     1146.52         573.26
 RESIDUAL      9      247.40          27.50
 TOTAL        11     1394.92

 FURTHER ANALYSIS OF VARIANCE
 SS EXPLAINED BY EACH VARIABLE WHEN ENTERED IN THE ORDER GIVEN

  DUE TO      DF          SS
 REGRESSION    2     1146.52
 C1            1     1110.54
 C3            1       35.93

              X1           Y    PRED. Y    ST.DEV.
 ROW          C1          C2      VALUE    PRED. Y    RESIDUAL     ST.RES.
  12        70.0       85.00      85.89       4.45       -0.89       -0.32 X

 X DENOTES AN OBS. WHOSE X VALUE GIVES IT LARGE INFLUENCE.

 DURBIN-WATSON STATISTIC = 2.72
 --
 --



                         PLOT OF W = K(1 - EXP(aX))   WHERE X IS TIME IN YEAR
```

```
SET C4
1.75 2.59 6.03 3.73 1.99 1.28 4.11 10.11 2.24 2.39 7.04 5.09 2.29
8.79 2.15 1.72 5.75 1.52 3.97 1.42
SET C5
1.17 2.42 3.49 2.14 1.63 0.89 3.29 4.26 1.33 0.73 4.06 3.34 1.15
3.45 1.67 0.83 2.25 1.31 4.12 1.51
LET C6=C5/C4
PLOT C5 VS C4
LOGE C4, C1
LOGE C5, C3
PLOT C3 VS C1
REGR C3 1 C1, C7, C8
PLOT C7 VS C8
MPLOT C3 VS C1 AND C8 VS C1
LET K1=0.7413
LET K2=2.101
LET K3=0.1035
LET K4=K1-K2*K3
LET K5=K1+K2*K3
PRINT K1 K4 K5
LET K1=-0.1675
LET K3=0.1349
LET K4=K1-K2*K3
LET K5=K1+K2*K3
PRINT K1 K4 K5
EXPO K1, K1
EXPO K4, K4
EXPO K5, K5
PRINT K1 K4 K5
GENE 1, 0.2, 13, C9
LET C10=0.9441*(C9**(0.7413-1))
MPLOT C6 VS C4 AND C10 VS C9
STOP
```

```
FILE: RATIO     OUTPUT   A1  VM/SP - CONVERSATIONAL MONITOR SYSTEM                                              PAGE 001

1MINITAB RELEASE 81.1 *** COPYRIGHT - PENN STATE UNIV. 1981
 MAY 10, 1985 *** NATIONAL UNIVERSITY OF SINGAPORE - LOCAL VERSION 02/12/1982
 STORAGE AVAILABLE   4800

    EXAMPLE OF LINEAR REGRESSION ANALYSIS FOR RATIO VARIABLES
                 THE HERMIT CRAB PROBLEM


 --
             PLOT OF WEIGHT OF SHELL (WS) VERSUS WEIGHT OF CRAB (WC)
 --
      C5
    4.50+
        -                                     *
        -            *            *
        -
        -
    3.60+
        -                 *       *      *
        -             *
        -
        -
    2.70+
        -
        -       *
        -            *       *
        -
    1.80+
        -     **
        -   *  *
        -   **
        -      *
    0.90+  * *  *
        -
        -
        -
        -
     0.0+
          +---------+---------+---------+---------+---------+C4
         1.0       3.5       6.0       8.5      11.0      13.5

 --
```

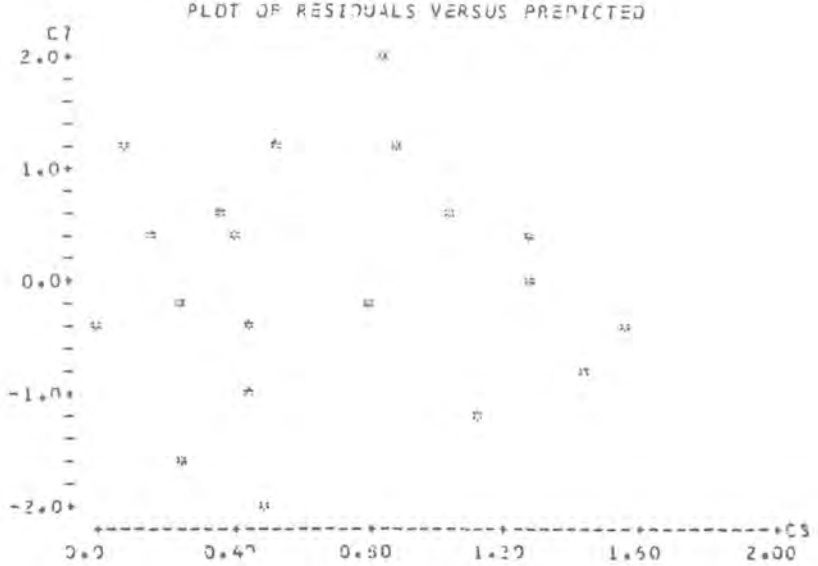
PLOT OF LOG(WS) VERSUS LOG (WC)



```
--
THE REGRESSION EQUATION IS
LOG(WS) = - 0.171 + 0.743 LOG(WC)

                                             ST. DEV.     T-RATIO =
        COLUMN       COEFFICIENT            OF COEF.     COEF/S.D.
        INTERCEPT      -0.1707               0.1353         -1.26
        SLOPE           0.7425               0.1038          7.15

THE ST. DEV. OF Y ABOUT REGRESSION LINE IS
S = 0.2899
WITH (  20-  2) =  18 DEGREES OF FREEDOM

R-SQUARED = 74.0 PERCENT
R-SQUARED = 72.5 PERCENT, ADJUSTED FOR D.F.

ANALYSIS OF VARIANCE

 DUE TO       DF          SS       MS=SS/DF
REGRESSION     1      4.29798       4.29798
RESIDUAL      18      1.51276       0.08404
TOTAL         19      5.81073
```

```
            X1           Y     PRED. Y    ST.DEV.
ROW         C1          C3       VALUE    PRED. Y    RESIDUAL      ST.RES.
  9       2.31      1.4493      1.5472     0.1377     -0.0979      -0.38 X
 19       1.35      1.4759      0.9342     0.0684      0.5417       2.06R

 R DENOTES AN OBS. WITH A LARGE ST. RES.
 X DENOTES AN OBS. WHOSE X VALUE GIVES IT LARGE INFLUENCE.

 DURBIN-WATSON STATISTIC = 1.68
```

--



```
--
      C3
   1.60+
       -                                               B
       -                            A           A   B  A
       -
       -                                        3
   1.20+                             A    A         A
       -                                    3
       -                                  B
       -
       -                    A        B
   0.80+                           29       A
       -
       -
       -                    B
       -               AA 2
   0.40+        A       9A
       -               3 A
       -         A  2
       -         3  A     A
       -        3
   0.0 +      B
       -      A           A
       -            A
       -
       -
  -0.40+
         +---------+---------+---------+---------+---------+C1
       0.0       0.50      1.00      1.50      2.00      2.50
--
--
   K1        0.741900
   K4        0.524347
   K5        0.959253
--
--
   K1       -0.160500
   K4       -0.452725
   K5        0.113925
--
   ANSWER =           0.8441
--
   ANSWER =           0.5359
--
   ANSWER =           1.1237
--
   K1        0.844087
   K4        0.635766
   K5         1.12357
--
--
--
```

5. ANALYSIS OF COVARIANCE

5.1 Introduction:

It originated as a procedure for improving an analysis of variance (to test for differences aong different groups, or treatments) by correcting for an uncontrollable variable which is influencing the variable of interest. Here we will consider it as a method for comparing bivariate relationships among different groups, or treatments. For those interested in a reference, the new (7th) edition of Snedecor and Cochran's "Statistical Methods" has good coverage (Chapter 18).

Let us illustrate the analysis of covariance model by an assigned problem based on Walford Plot data. The size measurements are lengths at consecutive annual winter rings in shells of clams living in the intertidal zone on Hudson Bay, Canada, at two tidal levels:

Om tide level

Lx : 2.7 3.0 3.2 3.3 3.4 3.6 3.9 4.3 4.3 4.8 5.6 5.7 6.0 6.6
6.9 7.8

Lx+1 : 5.8 6.3 5.4 5.1 6.1 6.4 5.9 7.9 8.1 7.5 8.9 8.5 9.2 9.0
8.8 10.2

Lx : 8.7 9.1 9.1 10.2 12.3

Lx+1 : 11.5 11.2 12.3 12.0 14.3

1.1 m tide level

Lx : 3.5 3.7 3.8 4.0 4.1 4.2 4.4 4.5 4.7 4.8 4.9 5.0
5.1 5.6 6.0 6.5

Lx+1 : 7.9 8.8 9.2 8.6 7.9 9.5 8.8 8.6 9.6 8.3 8.2 9.3
10.6 10.0 10.9 11.0

Lx : 7.3 7.7 9.0 9.4 112. 12.1

Lx+1 : 11.5 11.1 11.8 12.3 14.1 13.9

We wish to plot the data, check the plot for linearity, estimate the Walford Plot regreesion models and the corresponding Von Bertalanffy models, plot the model, and test whether the slopes and intercepts of the Walford Plot models differ between the two tidal levels.

5.2 Assigned problem:

1. Use the SET command to put Lx values into Cl, Lx+l values into C2, and a 0/1 code into C3 to represent tidal level.

2. Use the command "PLOT C2 Cl, C3" to produce a Walford Plot of the Lx+l versus Lx data, with points labeled by tidal level. Check that the trend, for each tidal level, is linear.

3. MINITAB does not have an "analysis of covariance" procedure, but we can do the necessary calculations using the REGR command. For 2 groups it is especially easy. What we want to do is regress Lx+l on 3 predictor variables: Lx in Cl, the tidal level code in C3, and the product of Lx times the tidal level code which we can put into C4 by 'MULT Cl BY C3, C4". Now do "REGR IN Cl C3 C4, ST. RESID. IN C5,PRED. Y IN C6."

4. If the regression coefficient for C4 is significant (check the tvalue) then the slopes of the Walford Plot models differ between tidal levels, and a test of intercepts has no meaning. Growth rate decreases with age faster at one tidal level than at the other. Asymptotic sizes probably, though not necessarily, differ.

5. If the slopes were not significantly different (they should be) then you would proceed to test for differences in intercepts, by seeing if the regression coefficient for C3 is significant. If the intercepts differ then you have different initial growth rates but the same relative decrease in growth rate with age. Asymptotic sizes will differ. If intercepts do not differ, then there is no evidence that the growth curves differ between tidal levels.

For a SAS analysis of covariance run you would use PROC GLM as follows:

```
TITLE ;
DATA COLDCLAMS;
INPUT YX YXP1 TL;
CARDS;

2.7     5.8     0
3.0     6.3     0
 ⋮       ⋮      ⋮
3.5     7.9     1
3.7     8.8     1
 ⋮       ⋮      ⋮
PROC GLM; CLASS TL;                 test of difference
MODEL YXP1 = TL YX YX*TL;           between slopes

PROC GLM; BY  TL;                   estimates separate
MODEL YXP1 = YX;                    slopes

PROC GLM; CLASS TL;                 estimates a common
MODEL YXP1 = TL YX;                 slope and tests for
                                    intercepts

PROC PLOT;                          does Walford Plot,
PLOT YXP1*YX = TL;                  with points coded
                                    by tidal level
```

For an analysis of covariance with more than 2 groups, the SAS package is probably the easiest.

5.4 Covariance analysis as an alternative to ratio variables

5.4.1 Assignment

Question: Do frogs differ in % water content between spring & fall? Since frogs sampled in spring and fall will probably be a mixture of sizes, one must also ask whether % water content varies with size of frog.

A typical approach: Let $Yi = \frac{\text{water wt.}}{\text{total wt.}}$ for frog i.

Collect n1 spring frogs and n2 fall frogs, determine water wt. and total wt, and calculate Yi for all n1 + n2 frogs. Test against the null hypothesis Ho:"that the mean Y is the same for spring frogs and fall frogs", using a t-test or a 1-way ANOVA. This is a *bad* appraoch - difficult to test what you want, difficult to interpret, and based on a ratio variable.

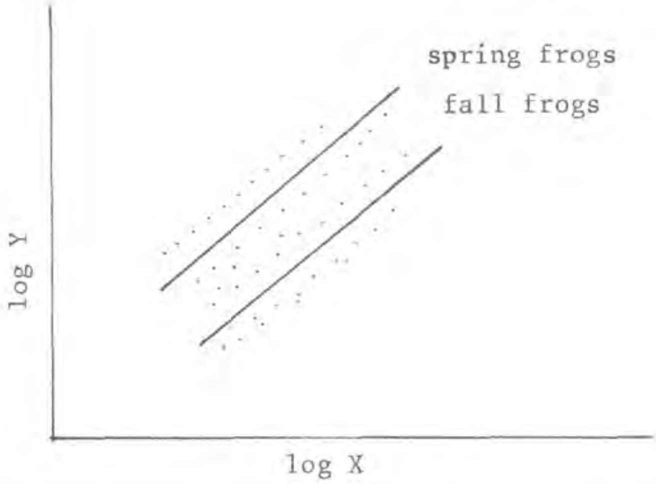
The covariance analysis approach: Let Y = water wt. and X = dry wt. Deal with the question of possible variation of "% water content" with size of frog first. The null hypothesis Ho is that as frog size varies the water wt. portion changes at the same % rate as does the dry wt. portion. For example, if a 10 gram frog is 6 grams water and 4 grams dry wt. then a 12 gram frog will be about 7.2 grams water (a 20% increase) and 4.8 grams dry weight (also a 20% increase), and the % water content is 60% in both cases. The model is $dY/Y = b\ dX/X$, and b=1 if Ho is true.

The nonlinear model is $Y = AX^b$ which is Y = AX *if* Ho is true. Thus Y/X = constant value A.

If b=1 then Y/X is not constant, but varies with size of frog.

If b<1 then big frogs have lower % water, and if b>1 they have higher % water. If spring and fall frogs of similar size have the same % water then A should be the same for both seasons.

The linear model is log Y = a + b log X (where a = log A). In covariance analysis one tests a sequence of hypotheses:



H1 : that the amount of variation about the regression lines is the same for both groups. If this is accepted, then

H2 : that the slopes (b-values) of the regression lines are the same. If this is accepted, then

H3 : that the common slope b is some specified value (e.g., b=1)

H4 : that the intercepts (a-values) of the two common-slope regression lines are the same.

In this situation these hypotheses have the following biological interpretations:

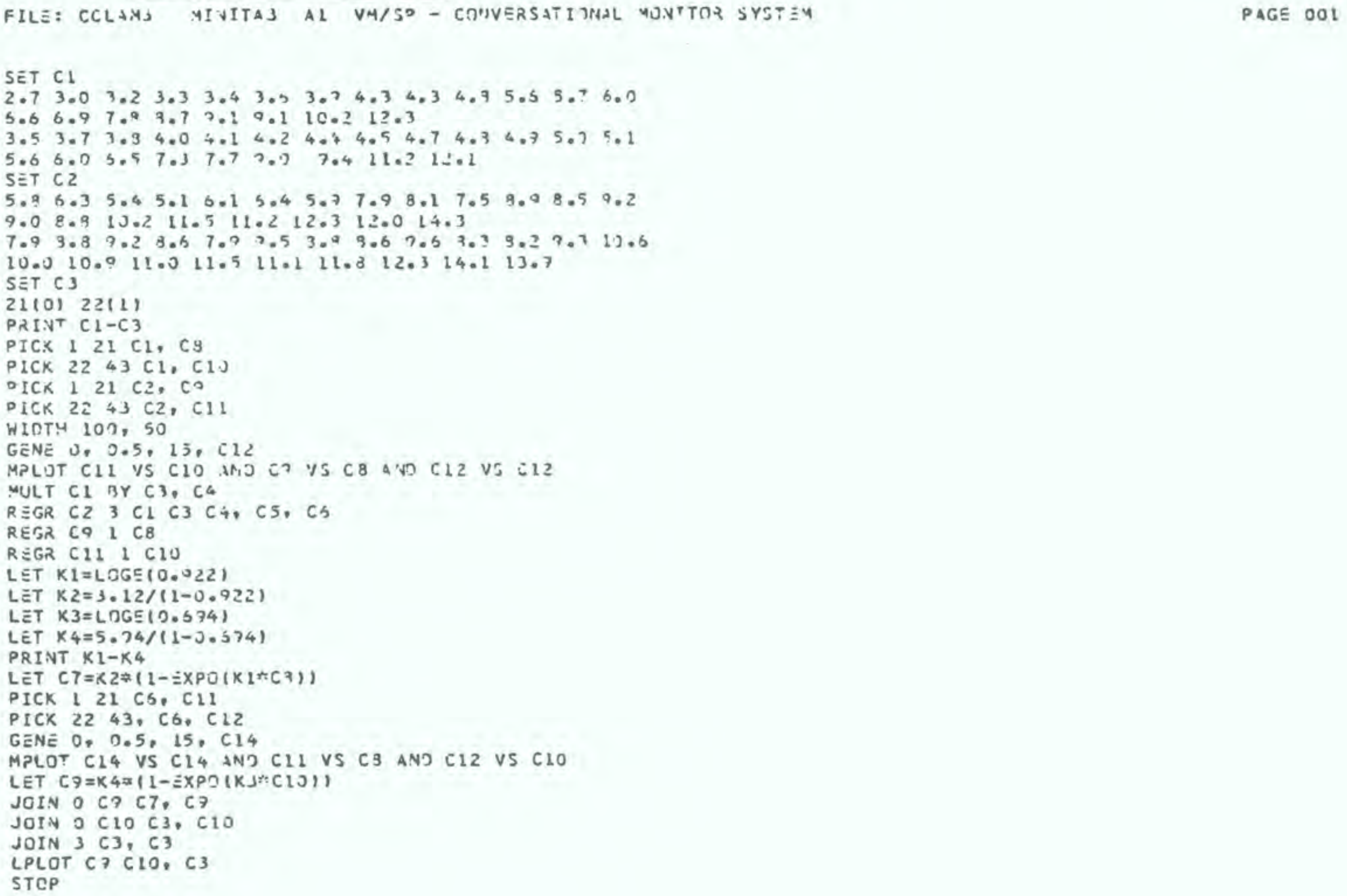
H1 : that variation in % water content among frogs of similar size does not differ between spring and fall.

H2 : that any variation (or lack thereof) in % water content with size of frog is similar in spring and fall. If this is accepted, then

H3 : for the common slope b=1, that % water content does not vary with size of frog.

H4 : that the average % water content of frogs of similar size does not differ between spring and fall.
Since a = log A, different a-values reflect different Y/X ratios.

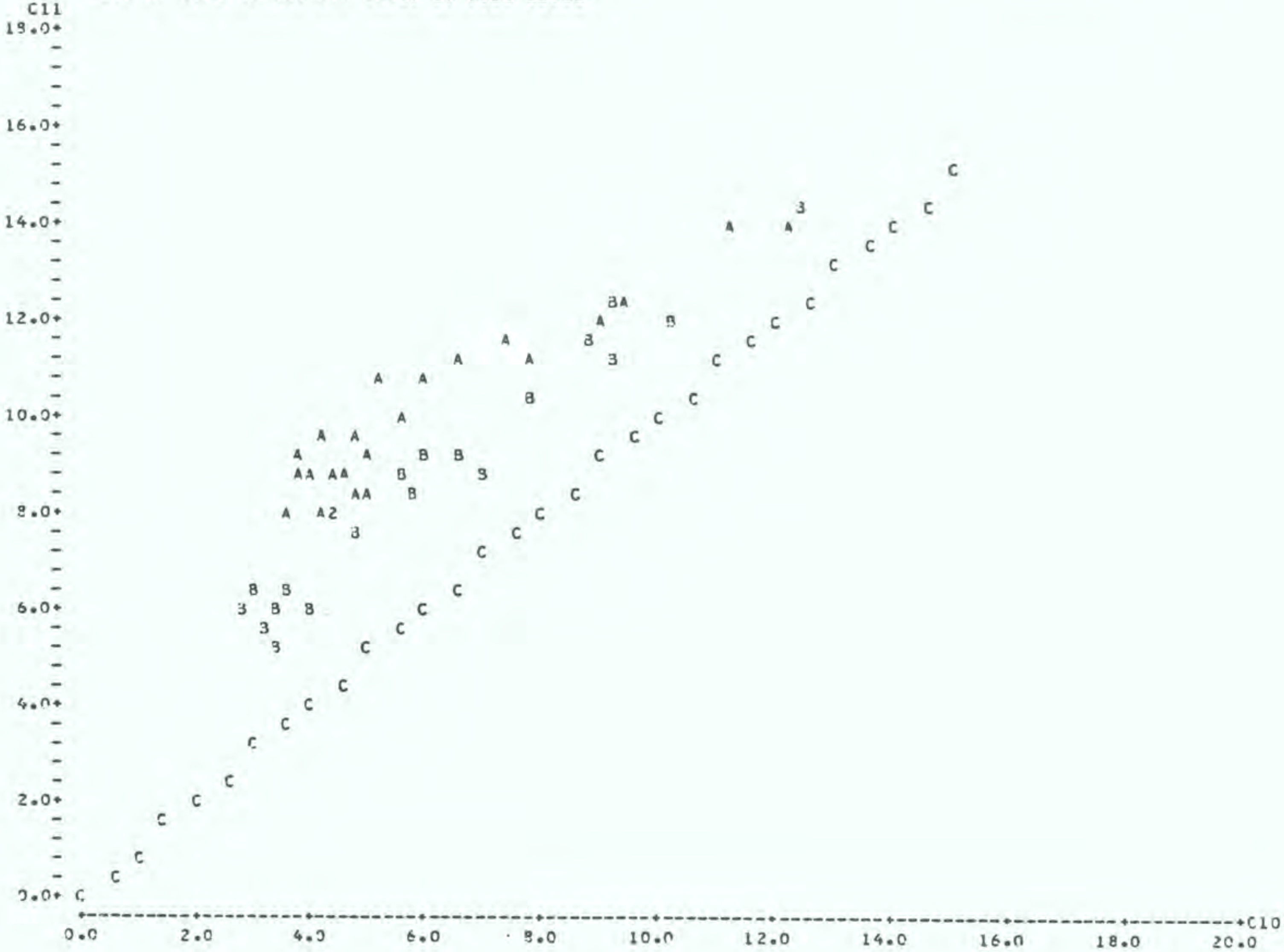
5.4.2. Job Listing and Output.



```
FILE: CCLAN3   MINITAB  A1  VM/SP - CONVERSATIONAL MONITOR SYSTEM                    PAGE 001


SET C1
2.7 3.0 3.2 3.3 3.4 3.5 3.9 4.3 4.3 4.9 5.6 5.7 6.0
6.6 6.9 7.8 8.7 9.1 9.1 10.2 12.3
3.5 3.7 3.8 4.0 4.1 4.2 4.4 4.5 4.7 4.8 4.9 5.0 5.1
5.6 6.0 6.5 7.0 7.7 9.0  9.4 11.2 12.1
SET C2
5.8 6.3 5.4 5.1 6.1 6.4 5.9 7.9 8.1 7.5 8.9 8.5 9.2
9.0 8.8 10.2 11.5 11.2 12.3 12.0 14.3
7.9 8.8 9.2 8.6 7.9 9.5 8.9 8.6 9.6 8.3 9.2 9.3 10.6
10.0 10.9 11.0 11.5 11.1 11.8 12.3 14.1 13.9
SET C3
21(0) 22(1)
PRINT C1-C3
PICK 1 21 C1, C8
PICK 22 43 C1, C10
PICK 1 21 C2, C9
PICK 22 43 C2, C11
WIDTH 100, 50
GENE 0, 0.5, 15, C12
MPLOT C11 VS C10 AND C9 VS C8 AND C12 VS C12
MULT C1 BY C3, C4
REGR C2 3 C1 C3 C4, C5, C6
REGR C9 1 C8
REGR C11 1 C10
LET K1=LOGE(0.922)
LET K2=3.12/(1-0.922)
LET K3=LOGE(0.694)
LET K4=5.94/(1-0.694)
PRINT K1-K4
LET C7=K2*(1-EXPO(K1*C8))
PICK 1 21 C6, C11
PICK 22 43, C6, C12
GENE 0, 0.5, 15, C14
MPLOT C14 VS C14 AND C11 VS C8 AND C12 VS C10
LET C9=K4*(1-EXPO(K3*C10))
JOIN 0 C9 C7, C9
JOIN 0 C10 C3, C10
JOIN 3 C3, C3
LPLOT C9 C10, C3
STOP
```

```
FILE: CCLAM3    OUTPUT    A1   VM/SP - CONVERSATIONAL MONITOR SYSTEM                                          PAGE 001

 --
 COLUMN        C1           C2          C3
 COUNT         43           43          43
 ROW         LXP1           LX          TL
    1       2.7000       5.8000         0.
    2       3.0000       6.3000         0.
    3       3.2000       5.4000         0.
    4       3.3000       5.1000         0.
    5       3.4000       5.1000         0.
    6       3.6000       6.4000         0.
    7       3.9000       5.9000         0.
    8       4.3000       7.9000         0.
    9       4.3000       8.1000         0.
   10       4.8000       7.5000         0.
   11       5.6000       8.9000         0.
   12       5.7000       8.5000         0.
   13       6.0000       9.2000         0.
   14       6.6000       9.0000         0.
   15       6.9000       8.8000         0.
   16       7.9000      10.2000         0.
   17       8.7000      11.5000         0.
   18       9.1000      11.2000         0.
   19       9.1000      12.3000         0.
   20      10.2000      12.0000         0.
   21      12.3000      14.3000         0.
   22       3.5000       7.9000         1.
   23       3.7000       8.8000         1.
   24       3.8000       9.2000         1.
   25       4.0000       8.6000         1.
   26       4.1000       7.9000         1.
   27       4.2000       9.5000         1.
   28       4.4000       8.8000         1.
   29       4.5000       8.6000         1.
   30       4.7000       9.6000         1.
   31       4.8000       8.3000         1.
   32       4.9000       8.2000         1.
   33       5.0000       9.3000         1.
   34       5.1000      10.6000         1.
   35       5.5000      10.0000         1.
   36       6.0000      10.9000         1.
   37       6.5000      11.0000         1.
   38       7.3000      11.5000         1.
   39       7.7000      11.1000         1.
   40       9.0000      11.8000         1.
   41       9.4000      12.3000         1.
   42      11.2000      14.1000         1.
   43      12.1000      13.9000         1.

 --
```

PLOT OF LXP1 VERSUS LX, CODED BY TIDE LEVEL

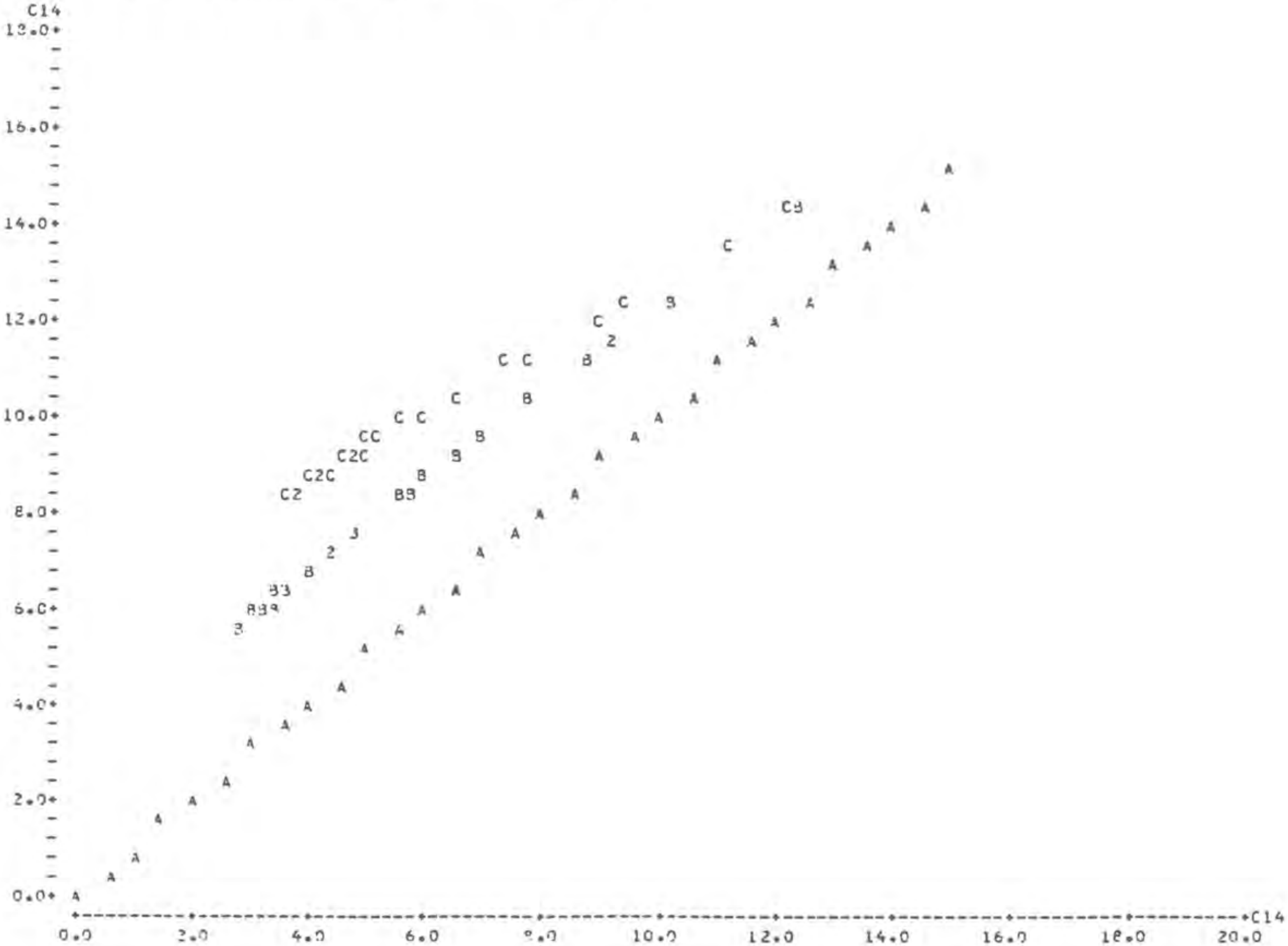


```
 THE REGRESSION EQUATION IS
 Y =     3.12 + 0.922 X1 +  2.81 X2
     - 0.228 X3

                                            ST. DEV.    T-RATIO =
        COLUMN      COEFFICIENT             OF COEF.    COEF/S.D.
        INTERCEPT       3.1222               0.3218         9.70
 X1     C1             0.92235              0.04955        18.62
 X2     C3              2.8133               0.4703         5.98
 X3     C4            -0.22791              0.07270        -3.13

 THE ST. DEV. OF Y ABOUT REGRESSION LINE IS
 S = 0.6019
 WITH (  43- 4) =  39 DEGREES OF FREEDOM

 R-SQUARED = 93.7 PERCENT
 R-SQUARED = 93.3 PERCENT, ADJUSTED FOR D.F.

 ANALYSIS OF VARIANCE

  DUE TO       DF          SS        MS=SS/DF
 REGRESSION     3     211.2982        70.4327
 RESIDUAL      39      14.1290         0.3623
 TOTAL         42     225.4272

 FURTHER ANALYSIS OF VARIANCE
 SS EXPLAINED BY EACH VARIABLE WHEN ENTERED IN THE ORDER GIVEN

  DUE TO       DF          SS
 REGRESSION     3     211.2982
 C1             1     184.9595
 C3             1      22.7785
 C4             1       3.5602

             X1          Y     PRED. Y    ST.DEV.
 ROW         C1         C2       VALUE    PRED. Y    RESIDUAL      ST.RES.
  21       12.3    14.3000     14.4672     0.3419     -0.1572      -0.34 X
  42       11.2    14.1000     13.7132     0.3061      0.3868       0.75 X
  43       12.1    13.9000     14.3382     0.3501     -0.4382      -0.90 X

 X DENOTES AN OBS. WHOSE X VALUE GIVES IT LARGE INFLUENCE.

 DURBIN-WATSON STATISTIC = 1.31
 —

 THE REGRESSION EQUATION (FOR TL=0 ONLY)
 Y =     3.12 + 0.922 X1
                                            ST. DEV.    T-RATIO =
        COLUMN      COEFFICIENT             OF COEF.    COEF/S.D.
        INTERCEPT       3.1222               0.3137         9.95
 X1     C8             0.92236              0.04830        19.10
```

```
THE ST. DEV. OF Y ABOUT REGRESSION LINE IS
S = 0.5867
WITH (  21- 2) =  19 DEGREES OF FREEDOM

R-SQUARED = 95.0 PERCENT
R-SQUARED = 94.8 PERCENT, ADJUSTED FOR D.F.

ANALYSIS OF VARIANCE

 DUE TO       DF            SS        MS=SS/DF
REGRESSION     1      125.5386        125.5386
RESIDUAL      19        6.5399          0.3442
TOTAL         20      132.0786

            X1             Y    PRED. Y     ST.DEV.
ROW         C8            C9      VALUE     PRED. Y    RESIDUAL      ST.RES.
 21       12.3        14.300     14.467       0.333      -0.167      -0.35 X

X DENOTES AN OBS. WHOSE X VALUE GIVES IT LARGE INFLUENCE.

DURBIN-WATSON STATISTIC = 1.89
--

THE REGRESSION EQUATION (FOR TL=1 ONLY)
Y =     5.94 + 0.694 X1
                                            ST. DEV.     T-RATIO =
       COLUMN       COEFFICIENT             OF COEF.     COEF/S.D.
       --                5.9355              0.3510         16.91
X1     C10              0.69444              0.05445        12.75

THE ST. DEV. OF Y ABOUT REGRESSION LINE IS
S = 0.6160
WITH (  22- 2) =  20 DEGREES OF FREEDOM

R-SQUARED = 89.1 PERCENT
R-SQUARED = 88.5 PERCENT, ADJUSTED FOR D.F.

ANALYSIS OF VARIANCE

 DUE TO       DF            SS        MS=SS/DF
REGRESSION     1       61.7174         61.7174
RESIDUAL      20        7.5390          0.3795
TOTAL         21       69.3064

DURBIN-WATSON STATISTIC = 1.72
--

  ESTIMATING THE GROWTH EQUATION (VON BERTALANFFY MODEL)
              L = K(1 - EXP(BX))
--
   K1      -0.0812101     B FOR TL = 0
   K2        40.0000      K FOR TL = 0
   K3      -0.365283      B FOR TL = 1
   K4        19.4119      K FOR TL = 1
--
```

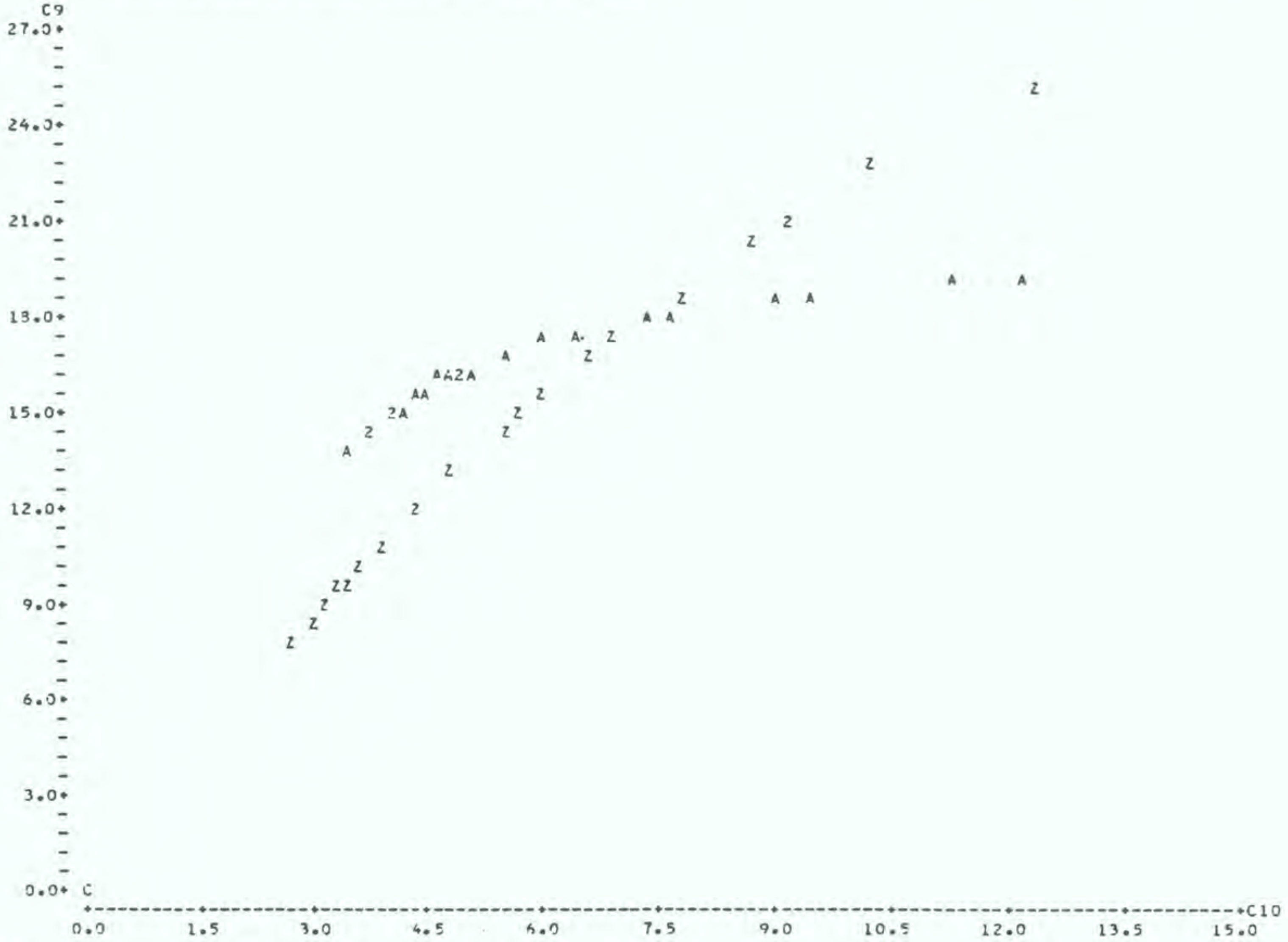
```
--
--

--
```

PLOT OF PREDICTED YXP1 VERSUS YX, CODED BY TL



```
FILE: CCLAM3   OUTPUT   A1  VM/SP - CONVERSATIONAL MONITOR SYSTEM                              PAGE 006
```

PLOT OF ESTIMATED GROWTH CURVES (VON BERTALANFFY MODEL)



```
*** MINITAB *** STATISTICS DEPT * PENN STATE UNIV. * RELEASE 81.1 *
STORAGE AVAILABLE     4900
```

```
FILE: ANCOVAR  SAS       A1  VM/SP - CONVERSATIONAL MONITOR SYSTEM                                      PAGE 001

TITLE EXAMPLE OF ANALYSIS OF COVARIANCE USING SAS;
DATA COLDCLAM;
INPUT YX YXP1 TL;
CARDS;
        27.000        59.000        0.0
        30.000        63.000        0.0
        32.000        54.000        0.0
        33.000        51.000        0.0
        34.000        61.000        0.0
        36.000        64.000        0.0
        39.000        57.000        0.0
        43.000        79.000        0.0
        43.000        81.000        0.0
        48.000        73.000        0.0
        56.000        89.000        0.0
        57.000        35.000        0.0
        60.000        92.000        0.0
        65.000        70.000        0.0
        69.000        88.000        0.0
        78.000       102.000        0.0
        87.000       115.000        0.0
        91.000       112.000        0.0
        91.000       123.000        0.0
       102.000       120.000        0.0
       123.000       143.000        0.0
        35.000        79.000        1.000
        37.000        88.000        1.000
        38.000        92.000        1.000
        40.000        86.000        1.000
        41.000        79.000        1.000
        42.000        95.000        1.000
        44.000        88.000        1.000
        45.000        36.000        1.000
        47.000        96.000        1.000
        48.000        83.000        1.000
        49.000        82.000        1.000
        50.000        93.000        1.000
        51.000       106.000        1.000
        56.000       100.000        1.000
        60.000       109.000        1.000
        65.000       110.000        1.000
        73.000       115.000        1.000
        77.000       111.000        1.000
        90.000       118.000        1.000
        94.000       123.000        1.000
       112.000       141.000        1.000
       121.000       139.000        1.000
PROC GLM; CLASS TL;
    MODEL YXP1 = TL YX YX*TL;
PROC GLM; BY TL;
    MODEL YXP1 = YX;
PROC GLM; CLASS TL;
    MODEL YXP1 = TL YX;
PROC PLOT;
    PLOT YXP1*YX = TL;
```

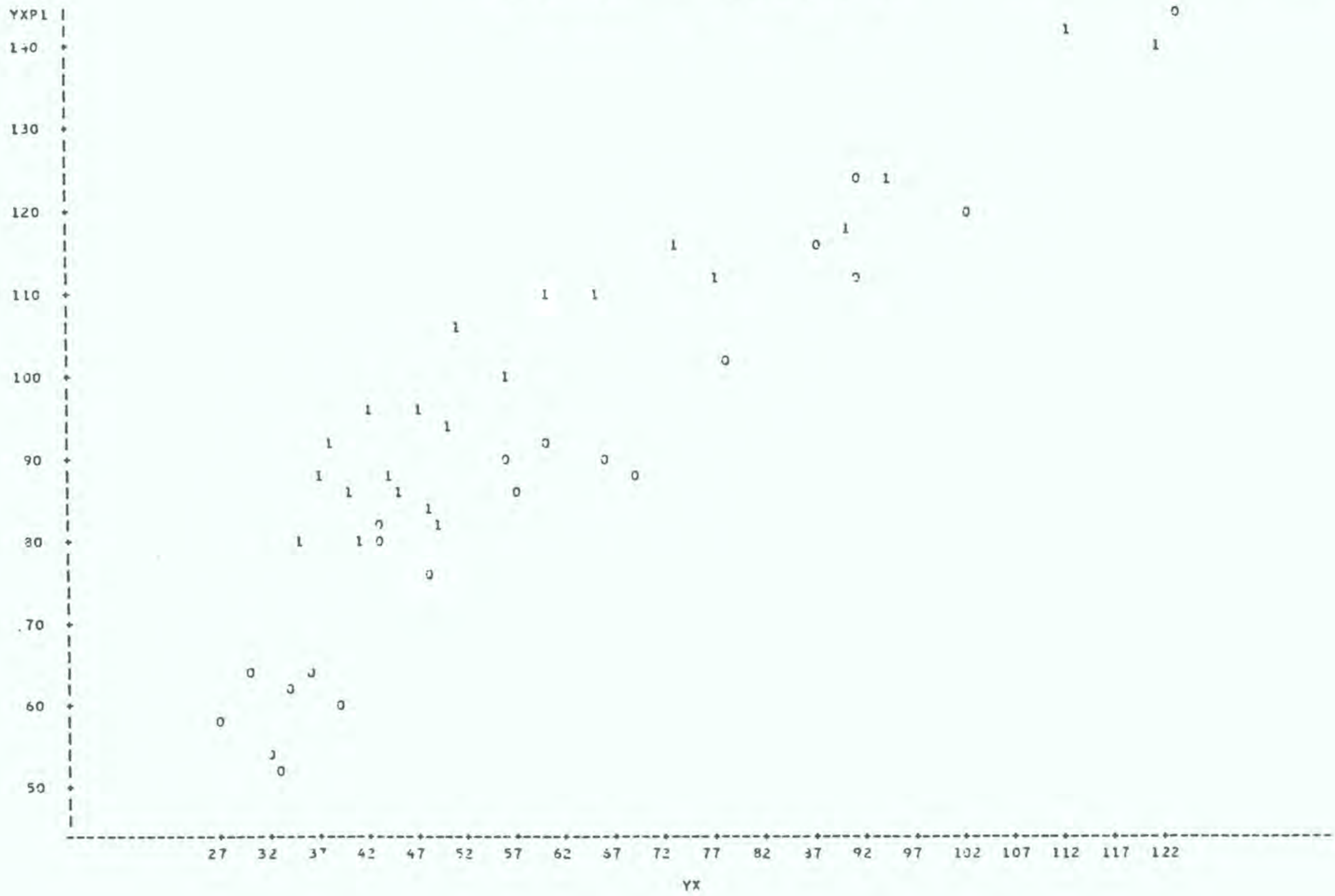
```
                                   EXAMPLE OF ANALYSIS OF COVARIANCE USING SAS            13:44 WEDNESDAY, MAY 8, 1985    2
                                      GENERAL LINEAR MODELS PROCEDURE

DEPENDENT VARIABLE: YXP1

SOURCE                  DF        SUM OF SQUARES          MEAN SQUARE       F VALUE           PR > F        R-SQUARE            C.V.
MODEL                    3        21129.70567092        7043.23522364        194.41           0.0001        0.937323          6.4334
ERROR                   39         1412.89898024          36.22817898                       ROOT MSE                      YXP1 MEAN
CORRECTED TOTAL         42        22542.60465116                                          6.01898488                     93.55813953

SOURCE                  DF           TYPE I SS       F VALUE      PR > F        DF         TYPE III SS       F VALUE         PR > F
TL                       1       2404.20421926         66.36      0.0001         1       1296.57011580         35.79         0.0001
YX                       1      18369.46698853        507.05      0.0001         1      17915.85492905        494.53         0.0001
YX*TL                    1        356.03446313          9.83      0.0033         1        356.03446313          9.83         0.0033
```

```
                                      EXAMPLE OF ANALYSIS OF COVARIANCE USING SAS          13:44 WEDNESDAY, MAY 8, 1985     3
                                                       TL=0

                                          GENERAL LINEAR MODELS PROCEDURE

DEPENDENT VARIABLE: YXP1

SOURCE                 DF       SUM OF SQUARES         MEAN SQUARE        F VALUE         PR > F        R-SQUARE           C.V.
MODEL                   1       12553.81321554      12553.81321554         364.72         0.0001        0.950434         6.8296
ERROR                  19         653.99630827         34.42085833                       ROOT MSE                      YXP1 MEAN
CORRECTED TOTAL        20       13207.80952381                                         5.86692921                    85.90476190

SOURCE                 DF            TYPE I SS      F VALUE       PR > F         DF        TYPE III SS      F VALUE       PR > F
YX                      1       12553.81321554       364.72       0.0001          1     12553.81321554       364.72       0.0001

                                   T FOR H0:      PR > |T|            STD ERROR OF
PARAMETER            ESTIMATE     PARAMETER=0                           ESTIMATE

INTERCEPT         31.22213617            9.95       0.0001              3.13652418
YX                 0.92235754           19.10       0.0001              0.04823724
```

```
                                   EXAMPLE OF ANALYSIS OF COVARIANCE USING SAS            13:44 WEDNESDAY, MAY 8, 1985    4
                                                      TL=1

                                           GENERAL LINEAR MODELS PROCEDURE

DEPENDENT VARIABLE: YXP1

SOURCE                   DF        SUM OF SQUARES          MEAN SQUARE         F VALUE           PR > F        R-SQUARE          C.V.
MODEL                     1        6171.68823713        6171.68823713          162.65           0.0001        0.890500        6.1072
ERROR                    20         758.90267197          37.94513360                          ROOT MSE                     YXP1 MEAN
CORRECTED TOTAL          21        6930.59090909                                              6.15995214                  100.86363636

SOURCE                   DF           TYPE I SS        F VALUE        PR > F          DF          TYPE III SS        F VALUE        PR > F
YX                        1        6171.68823713        162.65        0.0001           1        6171.68823713         162.65        0.0001

                                     T FOR H0:      PR > |T|            STD ERROR OF
PARAMETER                ESTIMATE   PARAMETER=0                          ESTIMATE

INTERCEPT             59.35520498         16.91        0.0001            3.50963941
YX                     0.69443764         12.75        0.0001            0.05445143
```

```
EXAMPLE OF ANALYSIS OF COVARIANCE USING SAS              13:44 WEDNESDAY, MAY 8, 1985    5
     GENERAL LINEAR MODELS PROCEDURE

          CLASS LEVEL INFORMATION

          CLASS     LEVELS    VALUES

          TL             2    0 1

  NUMBER OF OBSERVATIONS IN DATA SET = 43
```

```
                                       EXAMPLE OF ANALYSIS OF COVARIANCE USING SAS          13:44 WEDNESDAY, MAY 8, 1985    6
                                           GENERAL LINEAR MODELS PROCEDURE

DEPENDENT VARIABLE: YXP1

                   DF        SUM OF SQUARES          MEAN SQUARE          F VALUE          PR > F        R-SQUARE          C.V.
                    2        20773.67120779       10386.83560390           234.87          0.0001        0.921529        7.1079
                   40         1768.93344337          44.22333603                         ROOT MSE                    YXP1 MEAN
                   42        22542.60465116                                            6.65006286                  93.55813953

                   DF             TYPE I SS      F VALUE      PR > F          DF          TYPE III SS       F VALUE        PR > F
                    1         2404.20421326        54.37      0.0001           1        2277.87274935         51.51        0.0001
                    1        18369.46699453       415.38      0.0001           1       18369.46698953        415.38        0.0001
                    EXAMPLE OF ANALYSIS OF COVARIANCE USING SAS              13:44 WEDNESDAY, MAY 8, 1985    7
                       PLOT OF YXPL*YX     SYMBOL IS VALUE OF TL
```



```
EXAMPLE OF ANALYSIS OF COVARIANCE USING SAS          17:39 WEDNESDAY, MAY 8, 1985    1
      GENERAL LINEAR MODELS PROCEDURE

          CLASS LEVEL INFORMATION

          CLASS      LEVELS    VALUES

          TL            2       0 1

  NUMBER OF OBSERVATIONS IN DATA SET = 22
```

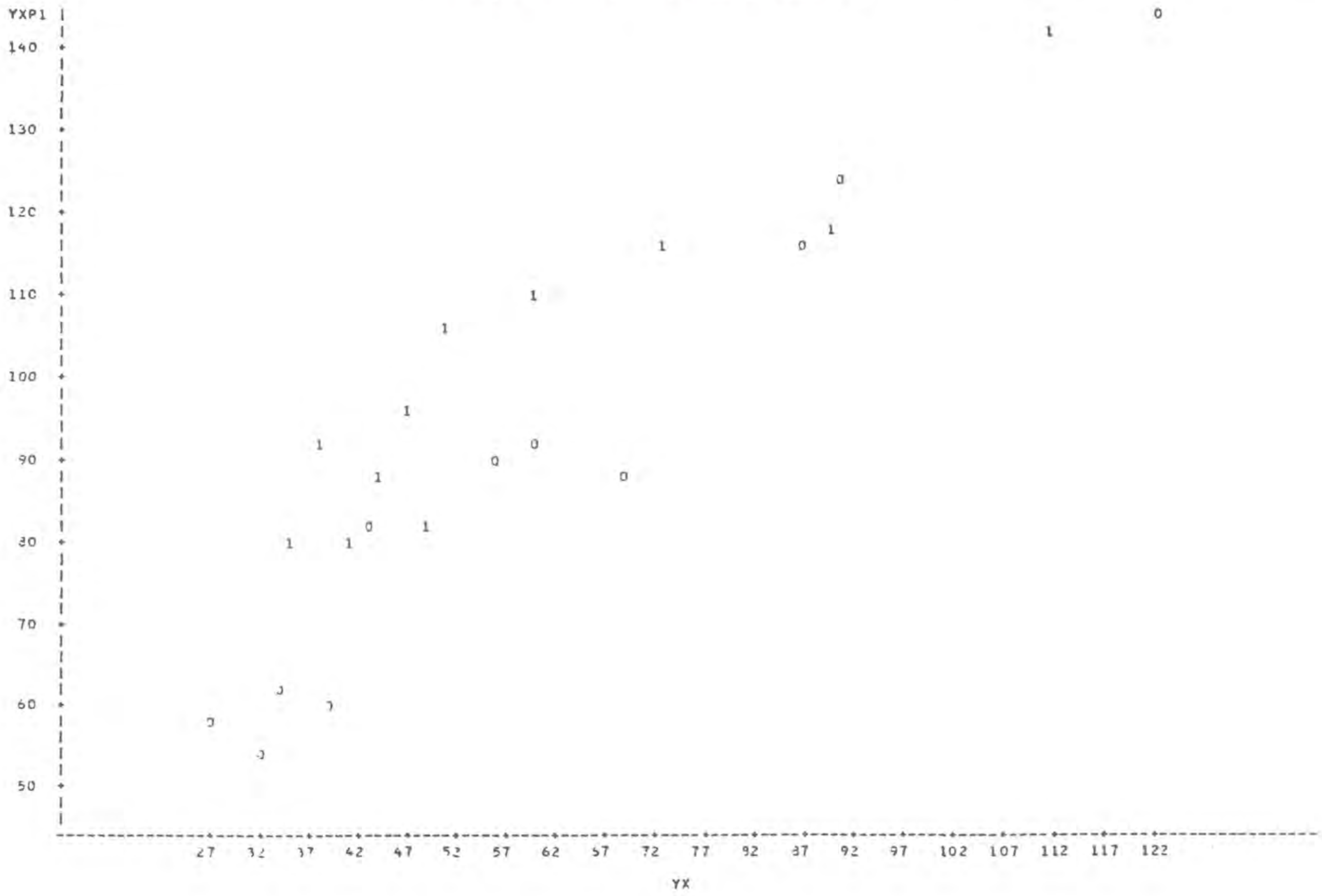
```
                                       EXAMPLE OF ANALYSIS OF COVARIANCE USING SAS           17:39 WEDNESDAY, MAY 8, 1985    2
                                          GENERAL LINEAR MODELS PROCEDURE

DEPENDENT VARIABLE: YXP1

SOURCE                    DF          SUM OF SQUARES           MEAN SQUARE           F VALUE              PR > F          R-SQUARE              C.V.
MODEL                      3          12457.39328909         4152.63276270             82.86              0.0001          0.932477            7.5312
ERROR                     18            902.10171191           50.11676177                             ROOT MSE                            YXP1 MEAN
CORRECTED TOTAL           21          13360.00000000                                                 7.07931930                          94.00000000

SOURCE                    DF               TYPE I SS       F VALUE       PR > F          DF              TYPE III SS        F VALUE         PR > F
TL                         1            916.54545455         18.29       0.0005           1             622.75396700          12.43         0.0024
YX                         1          11385.47459721        227.20       0.0001           1           10307.39191266         205.67         0.0001
YX*TL                      1            154.87823633          3.09       0.0958           1             154.87823633           3.09         0.0958
```

```
                                   EXAMPLE OF ANALYSIS OF COVARIANCE USING SAS           17:39 WEDNESDAY, MAY 8, 1985    3
                                                     TL=0

                                          GENERAL LINEAR MODELS PROCEDURE

DEPENDENT VARIABLE: YXP1

SOURCE                 DF       SUM OF SQUARES           MEAN SQUARE          F VALUE           PR > F          R-SQUARE             C.V.
MODEL                   1       8276.61566523          8276.61566523           189.97           0.0001          0.954767           7.5396
ERROR                   9        392.11160750            43.56795639                            ROOT MSE                        YXP1 MEAN
CORRECTED TOTAL        10       8563.72727273                                                6.60060273                       87.54545455

SOURCE                 DF          TYPE I SS       F VALUE        PR > F         DF          TYPE III SS        F VALUE          PR > F
YX                      1       8276.61566523       189.97        0.0001          1        8276.61566523         189.97          0.0001

                                  T FOR H0:        PR > |T|            STD ERROR OF
PARAMETER            ESTIMATE    PARAMETER=0                            ESTIMATE

INTERCEPT         30.03149400           6.50        0.0001             4.62311758
YX                 0.95711583          13.78        0.0001             0.06944192
```

```
                                        EXAMPLE OF ANALYSIS OF COVARIANCE USING SAS          17:39 WEDNESDAY, MAY 8, 1985    4
                                                          TL=1

                                            GENERAL LINEAR MODELS PROCEDURE

DEPENDENT VARIABLE: YXP1

SOURCE                   DF         SUM OF SQUARES          MEAN SQUARE         F VALUE          PR > F          R-SQUARE             C.V.
MODEL                     1         3264.73716832        3264.73716832           57.61          0.0001          0.864894           7.4936
ERROR                     9          509.99010441          56.66556716                        ROOT MSE                           YXP1 MEAN
CORRECTED TOTAL          10         3774.72727273                                           7.52765350                         100.45454545

SOURCE                   DF            TYPE I SS         F VALUE      PR > F          DF          TYPE III SS          F VALUE        PR > F
YX                        1         3264.73716832          57.61      0.0001           1        3264.73716832            57.61        0.0001

                                         T FOR H0:       PR > |T|              STD ERROR OF
PARAMETER                 ESTIMATE      PARAMETER=0                              ESTIMATE

INTERCEPT              56.92925043             9.23        0.0001               6.16710065
YX                      0.74809101             7.59        0.0001               0.09855756
```

```
                                            EXAMPLE OF ANALYSIS OF COVARIANCE USING SAS           17:39 WEDNESDAY, MAY 8, 1985    6
                                                  GENERAL LINEAR MODELS PROCEDURE

DEPENDENT VARIABLE: YXP1

SOURCE                    DF        SUM OF SQUARES          MEAN SQUARE       F VALUE          PR > F       R-SQUARE            C.V.
MODEL                      2        12303.02005176        6151.51002583        110.58          0.0001       0.920885          7.9347
ERROR                     19         1056.97994924          55.63052359                      ROOT MSE                      YXP1 MEAN
CORRECTED TOTAL           21        13360.00000000                                         7.45858724                    94.00000000

SOURCE                    DF             TYPE I SS     F VALUE     PR > F          DF          TYPE III SS      F VALUE       PR > F
TL                         1          916.54545455       16.48     0.0007           1        1167.55551974        20.99       0.0002
YX                         1        11386.47459721      204.68     0.0001           1       11386.47459721       204.68       0.0001
```

```
                    EXAMPLE OF ANALYSIS OF COVARIANCE USING SAS           17:39 WEDNESDAY, MAY 8, 1985    7
                       PLOT OF YXP1*YX    SYMBOL IS VALUE OF TL
```



```
COLUMN     DRYWT        WATER        SEASON (F=1,S=2)
COUNT         30           30           30
ROW
  1       2.1300       1.6500           1.
  2       3.0400       2.6300           1.
  3       4.4900       4.9700           1.
  4       2.3200       2.1000           1.
  5       1.5600       1.3000           1.
  6       4.4200       3.9900           1.
  7       8.3900       8.6000           1.
  8       5.5600       5.8300           1.
  9       4.4600       3.7100           1.
 10      11.4900      10.7100           1.
 11       6.7800       7.5700           1.
 12       3.1200       3.0500           1.
 13       9.5500       9.4400           1.
 14       8.7100       9.6100           1.
 15       1.3100       1.0300           1.
 16       2.8300       2.2200           2.
 17       2.9800       2.0400           2.
 18       3.1000       2.3600           2.
 19       2.0400       1.3100           2.
 20       1.5900       1.0500           2.
 21       1.9300       1.5000           2.
 22       1.5500       1.0700           2.
 23       4.7500       3.5800           2.
 24       8.4100       7.0800           2.
 25       3.1200       2.1300           2.
 26       2.2400       1.7700           2.
 27      18.4300      16.0300           2.
 28      11.2400       8.9300           2.
 29       2.6600       1.5900           2.
 30       1.2900       0.9200           2.

HERE IS THE "STATS ON A RATIO VARIABLE" APPROACH:

  LET 'PCTWAT'=100 ('WATER'/('WATER'+'DRYWT'))

COLUMN     DRYWT        WATER          SEASON     PCTWAT
COUNT         30           30             30          30
ROW
  1       2.1300       1.6500             1.     43.6508
  2       3.0400       2.6300             1.     46.3845
  3       4.4900       4.9700             1.     52.5370
  4       2.3200       2.1000             1.     47.5113
  5       1.5600       1.3000             1.     45.4545
  6       4.4200       3.9900             1.     47.4435
  7       8.3900       8.6000             1.     50.6180
  8       5.5600       5.8300             1.     51.1853
  9       4.4600       3.7100             1.     45.4100
 10      11.4900      10.7100             1.     48.2432
 11       6.7800       7.5700             1.     52.7526
 12      3.1200      3.0500         1.     49.4327
 13      9.5500      9.4400         1.     49.7104
 14      8.7100      9.6100         1.     52.4563
 15      1.3100      1.0300         1.     44.0171
 16      2.8100      2.2200         2.     43.5604
 17      2.9900      2.0400         2.     40.6375
 18      3.1000      2.3600         2.     43.2234
 19      2.0400      1.3100         2.     39.1045
 20      1.5900      1.0500         2.     35.7727
 21      1.9300      1.5000         2.     43.7318
 22      1.5500      1.0700         2.     40.8397
 23      4.7500      3.5800         2.     42.9772
 24      8.4100      7.0800         2.     45.7069
 25      3.1200      2.1300         2.     40.5714
 26      2.2400      1.7700         2.     44.1397
 27     19.4300     16.0100         2.     46.5177
 28     11.2400      8.9300         2.     44.2737
 29      2.6600      1.5900         2.     37.4118
 30      1.2900      0.9200         2.     41.6290

-- ONEWAY ANALYSIS OF VARIANCE, ON 'PCTWAT', BETWEEN SEASONS

ANALYSIS OF VARIANCE

DUE TO        DF        SS     MS=SS/DF     F-RATIO

SEASON         1     284.04     284.04       35.39
ERROR         28     224.73       8.03
TOTAL         29     508.76

SEASON        N          MEAN       ST. DEV.

FALL         15         48.45         3.08
SPRING       15         42.30         2.56

INDIVIDUAL 95 PERCENT C. I. FOR LEVEL MEANS
(BASED ON POOLED STANDARD DEVIATION)
         +---------+---------+---------+---------+---------+---------+
FALL                                          I-------I*-------I
SPRING       I------I-------I
         +---------+---------+---------+---------+---------+---------+
       40.0      42.0      44.0      46.0      48.0      50.0      52.0

NOW HERE IS THE "LOG-LOG ANALYSIS OF COVARIANCE" APPROACH:

-- LET 'DRYWT'=LOGE('DRYWT')
-- LET 'WATER'=LOGE('WATER')

COLUMN   LN(DRYWT)   LN(WATER)      SEASON
COUNT          30          30          30
ROW
```

```
   1      0.75612     0.50078         1.
   2      1.11196     0.95698         1.
   3      1.50185     1.60342         1.
   4      0.94157     0.74194         1.
   5      0.44469     0.26236         1.
   6      1.48614     1.38377         1.
   7      2.12704     2.15176         1.
   8      1.71560     1.76302         1.
   9      1.49515     1.31103         1.
  10      2.44148     2.37118         1.
  11      1.91579     2.02419         1.
  12      1.13733     1.11514         1.
  13      2.25654     2.24496         1.
  14      2.16447     2.26290         1.
  15      0.27003     0.02956         1.
  16      1.04128     0.79751         2.
  17      1.09192     0.71295         2.
  18      1.13140     0.85966         2.
  19      0.71295     0.27003         2.
  20      0.46373     0.04379         2.
  21      0.65752     0.40547         2.
  22      0.43825     0.06766         2.
  23      1.55314     1.27536         2.
  24      2.12942     1.95727         2.
  25      1.13783     0.75612         2.
  26      0.90549     0.57098         2.
  27      2.91398     2.77446         2.
  28      2.41743     2.15942         2.
  29      0.97333     0.46373         2.
  30      0.25464    -0.08338         2.

 LN(WATER)
     2.80+                                            B
         -
         -
         -                                    A
         -                                AA  B
     2.10+                               A
         -                            A  B
         -                          A
         -
         -                      A
     1.40+                      A
         -                      AB
         -                 A
         -                 A
         -                BB
     0.70+             A   2
         -            AB
         -          0    B
         -       A   B
         -
     0.00+     2 BB
```

```
                    +---------+---------+---------+---------+---------+
                  0.00      0.70      1.40      2.10      2.80      3.50
                                                              LN(DRYWT)

- LET 'DRWT.SN'='LN(DRYWT)'*'SEASON'

COLUMN  LN(DRYWT)  LN(WATER)        SEASON     DRWT.SN
COUNT          30         30            30          30
ROW
   1      0.75612    0.50078            1.     0.75612
   2      1.11186    0.96693            1.     1.11186
   3      1.50185    1.60342            1.     1.50185
   4      0.84157    0.74194            1.     0.84157
   5      0.44469    0.26236            1.     0.44469
   6      1.48614    1.38379            1.     1.48614
   7      2.12704    2.15176            1.     2.12704
   8      1.71560    1.76302            1.     1.71560
   9      1.49515    1.31103            1.     1.49515
  10      2.44148    2.37118            1.     2.44148
  11      1.91399    2.02419            1.     1.91398
  12      1.13783    1.11514            1.     1.13783
  13      2.25654    2.24496            1.     2.25654
  14      2.16447    2.26280            1.     2.16447
  15      0.27003    0.02956            1.     0.27003
  16      1.04028    0.79751            2.     2.08055
  17      1.09192    0.71295            2.     2.18385
  18      1.13140    0.85866            2.     2.26280
  19      0.71295    0.27003            2.     1.42590
  20      0.46373    0.04879            2.     0.92747
  21      0.65752    0.40547            2.     1.31504
  22      0.43825    0.06766            2.     0.87651
  23      1.55814    1.27536            2.     3.11629
  24      2.12942    1.95727            2.     4.25884
  25      1.13783    0.75612            2.     2.27567
  26      0.80648    0.57098            2.     1.61295
  27      2.91398    2.77446            2.     5.82796
  28      2.41948    2.18942            2.     4.83896
  29      0.97933    0.46373            2.     1.95665
  30      0.25464   -0.08339            2.     0.50928


-- REGRESS 'LN(WATER)' ON 3 PREDICTORS 'LN(DRYWT)' 'SEASON' 'DRWT.SN'

                                              ST. DEV.    T-RATIO =
       COLUMN       COEFFICIENT               OF COEF.    COEF/S.D.
       --              -0.0811                0.1180
X1     LN(DRYWT)        1.16970               0.07586
X2     SEASON          -0.16729               0.06958
X3     DRWT.SN         -0.04067               0.04625        -0.88  (P>0.05) NS

CONCLUSION: NO SLOPE DIFFERENCE BETWEEN SEASONS.
               WHICH MEANS THAT IF PERCENT WATER VARIES WITH
               SIZE OF FROG (I.E., WITH DRY WT.) THEN IT
```

```
             DOES SO IN THE SAME MANNER IN BOTH SEASONS.


- REGRESS 'LN(WATER)' ON 2 PREDICTORS 'LN(DRYWT)' 'SEASON'

                                          ST. DEV.          T-RATIO =
       COLUMN      COEFFICIENT            OF COEF.          COEF/S.D.
       --              0.00609            0.06366
X1     LN(DRYWT)       1.10608            0.02284  (RE. B=1)  4.64  (P<0.01) **
X2     SEASON         -0.22140            0.03239            -6.33  (P<0.01) **

R-SQUARED = 97.0 PERCENT

CONCLUSIONS: (1) FOR A GIVEN SIZE (I.E., DRY WEIGHT) FROG,
                 WATER CONTENT DIFFERS BETWEEN SEASONS.
             (2) SINCE B>1, PERCENT WATER CONTENT INCREASES
                 WITH SIZE (I.E., DRY WEIGHT) OF FROG.


FALL:    LN(WATER) =  0.00609 + 1.10608*LN(DRYWT) - 0.2214*(1)
                   = -0.21531 + 1.10608*LN(DRYWT)
             WATER = 0.8063*(DRYWT**1.10608)

SPRING: LN(WATER) =  0.00609 + 1.10608*LN(DRYWT) - 0.2214*(2)
                  = -0.43671 + 1.10608*LN(DRYWT)
            WATER = 0.6462*(DRYWT**1.10608)


RLINE LN(WATER), LN(DRYWT)     \  ROBUST REGRESSION ESTIMATES SLOPE VERY
SLOPE =      1.1062            /     CLOSE TO SAME VALUE AS MODEL I REGRESSION.


PREDICTED LN(WATER)
               C6
            2.80+                                           B
                -      A=FALL
                -      B=SPRING                     A
                -
                -                              AA   B
            2.10+                              A
                -                           A  B
                -
                -                         A
                -
            1.40+                     3
                -                      B
                -
                -                5
                -                3
            0.70+             A BB
                -            A
                -           B B
                -       A  U
                -     A  B
```

```
              7.30+     B B
                    +---------+---------+---------+---------+---------+
                  0.00      0.70      1.40      2.10      2.80      3.50
                                                           LN(DRYWT)

  FALL:
    PRED. %WATER = 100*0.8063*(DRYWT**1.10608)/(DRYWT + PRED. WATER)
                 = 100*0.8063*(DRYWT**1.10608)/(DRYWT + 0.8063*(DRYWT**1.10608))
  SPRING:
    PRED. %WATER = 100*0.6462*(DRYWT**1.10608)/(DRYWT + PRED. WATER)
                 = 100*0.6462*(DRYWT**1.10608)/(DRYWT + 0.6462*(DRYWT**1.10608))

    PREDICTED WATER
               C14
              51.0+                              A
                  -                      AA A
                  -                  A
                  -               A
                  -            3
              48.0+
                  -         2
                  -      AA                                       B
                  -
                  -    AA                       B
              45.0+                      B
                  -
                  -
                  -             B
                  -                                A=FALL
              42.0+        32                      B=SPRING
                  -       B
                  -      2
                  -     2
                  -    B
              39.0+
                    +---------+---------+---------+---------+---------+
                   0.0       4.0       8.0      12.0      16.0      20.0
                                                               DRYWT

           %WATER
              56.0+
                  -
                  -
                  -
                  -            A     A    A
              52.0+
                  -               A
                  -                      A
                  -         A               A
                  -
              48.0+                              A
```

```
        -       A    A
        -         A                                     B
        -    A      A         B
        -
    44.0+    A 2BB                    B
        -         3   B
        -
        -    B
        -     B  BB
    40.0+     B                         A=FALL
        -      B                        B=SPRING
        -
        -        3
        -
    36.0+
          +---------+---------+---------+---------+---------+
         0.0       4.0       8.0     12.0      16.0      20.0
                                                     DRYWT

  %WATER    FALL FROGS
    53.5+   ----------
        -                  A
        -            A          A
        -
        -
    51.0+               A              B
        -                      2B B
        -
        -         A        B      A
        -               B
    48.5+            3
        -                              A
        -       A 2  A
        -       B
        -      B  A
    46.0+     B               A=%WATER
        -    BA      A        B=PRED %WATER
        -
        -
        -    A
    43.5+      A
          +---------+---------+---------+---------+---------+
         0.0       4.0       8.0     12.0      16.0      20.0
                                                     DRYWT

  %WATER    SPRING FROGS
            ------------
    47.0+                                               B
        -                                               A
        -
        -                      A      B
        -                      B
```

```
      44.5+                             A
          -       AA
          -      A      B
          -         A   A
          -
      42.0+        32
          -    A  B
          -     A2
          -     2  AA
          -    DA                   A=%WATER
      39.5+                         B=PRED %WATER
          -      A
          -
          -
          -        A
      37.0+
            +---------+---------+---------+---------+---------+
           0.0       4.0       8.0      12.0      16.0      20.0
                                                       DRYWT
```

6. INTODUCTION TO MULTIVARIATE ANALYSIS

6.1 A priori structure in a data set:

In general a data set has n observations (usually n samples) on p variables. Typically there are n rows and p columns, so the data matrix can be represented by

$$X = \begin{array}{cc} & p \\ n & \left[\begin{array}{ccc} \quad & \quad & \quad \\ & & \\ & & \\ & & \end{array}\right] \end{array}$$

Quite often the data matrix has *a priori* structure. That is, we perceive the rows and/or the columns to fall into groups which existed conceptually *before* we examined the collected data, and preferably before we collected the data. In fact, this *a priori* structure usualy represents the design of the data analysis which will be applied.

The figure 6.2 shows the various types of *a priori* structure of a data matrix. The dashed horizontal or vertical lines represent partitions of rows or columns into groups of rows or of columns. Each example suggests a category of types of data analysis, and the univariate cases should be familiar to you. (Figure taken from Green (1979) with permission).

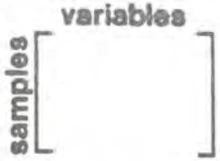
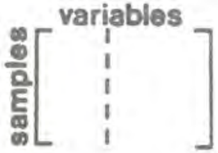
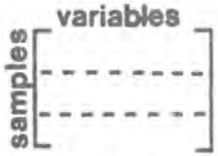
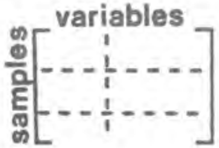
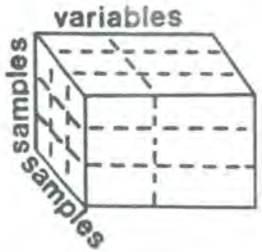
PARTITIONING	SYMBOLIC	MULTIVARIATE MODELS	UNIVARIATE MODELS
NONE		principal components analysis factor analysis cluster analysis	a priori restriction to one factor or component
VARIABLES		canonical correlation analysis	multiple regression and correlation
SAMPLES		MV analysis of variance and discriminant analysis	analysis of variance
VARIABLES AND SAMPLES		MV analysis of covariance and discriminant analysis with covariance	analysis of covariance
MULTI-DIMENSIONAL		factorial MV analysis of variance and covariance designs	factorial analysis of variance and covariance

FIGURE 6.2. Data matrices and statistical models

6.3 Types of data matrices and statistical analyses

6.3.1 No a priori structure:

We have simply collected an observational data set, n observations on p variables. The p variables are not divided into "predicted" and "predictor" types, and the n observations are not divided into *a priori* groups (such as different treatments, locations, times). If p = 1 we have the univariate case, and we would usually summarize the data graphically or do summary statistics appropriate for one column of data. These would be sample statistics such as x, s^2, s, SE and .95 cl on x. If p>1 we can of course do this for each variable, but besides looking at the pattern of variation of each variable we can also look at the pattern of covariation between each pair of variables. We speak of the "mean vector" and the "deviation cross-prodicts matrix" and the "variance - covariance matrix".

For p = 3 these are:

$$\bar{X} = \begin{bmatrix} \bar{x}_1 & \bar{x}_2 & \bar{x}_3 \end{bmatrix}$$

$$W = \begin{bmatrix} \Sigma(x_1-\bar{x}_1)^2 & \Sigma(x_1-\bar{x}_1)(x_2-\bar{x}_2) & \Sigma(x_1-\bar{x}_1)(x_3-\bar{x}_3) \\ \Sigma(x_2-\bar{x}_2)(x_1-\bar{x}_1) & \Sigma(x_2-\bar{x}_2)^2 & \Sigma(x_2-\bar{x}_2)(x_3-\bar{x}_3) \\ \Sigma(x_3-\bar{x}_3)(x_1-\bar{x}_1) & \Sigma(x_3-\bar{x}_3)(x_2-\bar{x}_2) & \Sigma(x_3-\bar{x}_3)^2 \end{bmatrix}$$

$$D = W/(n-1) = \begin{bmatrix} s_1^2 & s_{12} & s_{13} \\ s_{21} & s_2^2 & s_{23} \\ s_{31} & s_{32} & s_3^2 \end{bmatrix} \quad \begin{bmatrix} S_1^2 & S_{12} & S_{13} \\ & S_2 & S_{23} \\ & & S_3^2 \end{bmatrix}$$

If the data were standardized $\left[(x - \bar{x}J)/sJ\right]$ we would have the correlation matrix

$$R = \begin{bmatrix} 1 & r_{12} & r_{13} \\ r & 1 & r_{23} \\ r_{31} & r_{32} & 1 \end{bmatrix} = \begin{bmatrix} 1 & r_{12} & r_{13} \\ & 1 & r_{23} \\ & & r \\ & & 1 \end{bmatrix} = \begin{bmatrix} r_{12} & r_{13} \\ & r_{23} \\ & \end{bmatrix}$$

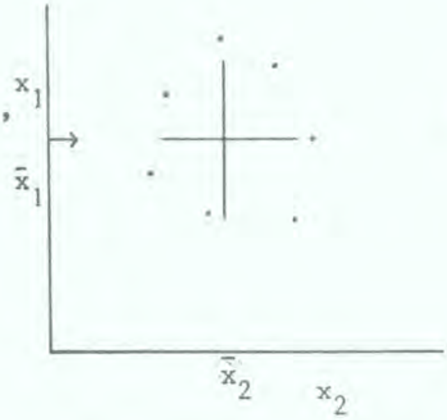
In the univariate case, these are all quite trivial and boring:

$$\bar{x} = [\,\bar{x}\,] \text{ , } D = [S^2] \text{and } R = [1]$$

In the multivariate case we will:

(1) locate the mean vector of the data in a p-dimensional space,

e.g.,



for p = 2;

(2) test the D or the R matrix for "structure" that is, against the null hypothesis that

$$R \text{ estimates } \rho = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} = \text{an identity matrix.}$$

(all true correlations are zero)

A confidence bound on a set of standardized observations, which yielded an R matrix that was an identify matrix, would be a sphere (for p = 3, a circle for p = 2, a spheroid for p > 3).

Thus the usual test against Ho = "identity matrix" is called a test of sphericity, or the "sphericity test". If "sphericity" is rejected, in favor of an elliptical shape then one is saying that <u>at least some</u> correlations appear to be non-zero, which is saying that the data "have structure";

(3) describe any structure, given that the Ho : "no structure" has been rejected. Some of the methods consist of identifying which variables are correlated. Other methods consist of finding, <u>a posteriori</u>, the partitions - of variables (vertical) or of observations (horizontal) - which "best" describe the structure in the data. Of course one can not turn around and test the "significance" of such partitions. It would make as much sense to separate a group of people into those shorter than the median height and those taller, and then test whether height differed between the groups. The only valid test, of "structure" versus "no structure", has already been done. Cluster analysis is a category of methods for partitioning observations into sets. If the n-by-p data matrix is partitioned so that the n observations are grouped into g<n sets of observations, the we have a cluster analysis solution which has reduced the data matrix from n-by-p to g-by-p. If it is a "best" cluster analysis solution then it has in some sense done so in a manner that has retained the maximum possible information about the structure in the original unreduced data matrix. Analysis by ordination similarly reduces the data matrix so as to retain maximum possible information in the reduced description, but it does so by partitioning the variables into "best" sets. Thus the original p-dimensional space is reduced to k<p - dimensional space. Again, tests of the "significance" of the partitioning are not very meaningful.

6.3.2 <u>A priori</u> structure is partitioning of variables into groups:

Let us deal with the simplest and most common case where one partition separates the variables into two sets, one on the left which is the criterion set and one on the right which is the predictor set. Here the a priori structure tells us that the data were collected in order to predict the left-hand variables from the right-hand variables. If a linear additive model is applicable (perhaps after transformation of the original variables) then we have the general linear model, or if neither set is clearly the predictor or the criterion variable set, then we have canonical correlation analysis. In the univariate case (one variable in the left-hand set), we have multiple regression and multiple correlation analysis, respectively. If, in the univariate case, the right-hand set also contains only one variable, then we have simple linear regression and correlation, respectively.

6.3.3 A priori structure is partitioning of observations into groups:

Observations can be partitioned into any number of sets. They may be treatments, locations, times, or combinations of those, but this a priori structure tells us that the data were collected in order to predict values on the p variables from knowledge of the group membership of an observation (or vice versa). Of course a test of whether there is any predictive power (whether the groups in fact differ on the variables) is the first step, and in the univariate case (one column in the data matrix) that is the only possible step (an ANOVA). When $p > 1$, and the test for group differences on the variables is significant, we would usually proceed to describe the group differences in terms of the relative contributions by the different variables to the group differences. The test would be a MANOVA (an acronym with obvious meaning). The descriptive analysis goes by various names: discriminant analysis, multiple discriminent analysis, and canonical analysis.

6.3.4 A priori structure which is a combination, or a multiple, of the above:

If partitioning is both vertical and horizontal then there is a predictor set of variables used to predict a criterion set of variables, but the observations on all these variables fall into dfferent _a priori_ groups. This is univariate (one criterion variable) or multivariate (>1 criterion variable) analysis of covariance. If groups, or treatment levels, are defined for more than one factor then we have a factorial UV or MV analysis of variance or covariance. Any univariate linear additive model - any regression, ANOVA or covariance design in existence - is just a special case of multivariate model.

6.4 Example of some basic calculations for multivariate analysis

6.4.1 Description

You have learned the necessary calculations and how to do them in MINITAB and in APL; matrix addition, multiplication, inversion, and transposition, and finding roots and vectors.

MINITAB does not have a command to calculate a W matrix or a D matrix, but the MINITAB job file (section)shows how to do it.

Enter these 3-variable data into C1 - C3:

$$\begin{bmatrix} 4.5 & 2.9 & 3.0 \\ 4.9 & 4.1 & 3.1 \\ 4.2 & 3.5 & 3.3 \\ 4.1 & 3.8 & 2.9 \\ 4.7 & 3.6 & 3.6 \\ 4.4 & 3.7 & 3.5 \end{bmatrix}$$

Obviously there are n=6 observations on p=3 variables.
Calculate the mean vector by doing

```
AVER C1, K1
AVER C2, K2
AVER C3, K3
```

Now use the MINITAB job on the attached sheet to calculate W and D. You should find that

$$D = \begin{bmatrix} 0.0907 & 0.0280 & 0.0213 \\ 0.0280 & 0.1600 & 0.0100 \\ 0.0213 & 0.0100 & 0.0787 \end{bmatrix}$$

Repeat this job run, but this time change lines 7-9, of the job file so the data are standardized on each variable. (Change to: LET Ci=(Ci-AVER(Ci))/STAN(Ci). Now D wil be the R matrix. Is it the same as you obtain by doing "CORR C1-C3, M1"?
Finally, with M1 containing the R matrix, do

```
EIGEN   M1, C4, M2
PRINT   C4
PRINT   M2
```

C4 contains the eigenvalues, which sum to 3 as did the diagonal of R. The columns of M2 contain the eigenvector coefficients associated with the eigenvalue above it. You have just done a principal components analysis!

6.4.2 Program for calculating W and D : MINITAB

```
1  NRAND 50, 10, 2, C1
2  NRAND 50, 12, 3, C2
3  NRAND 50, 15, 4, C3
4  SET C4
5  3(49)
6  LET C4=1/C4
7  LET C1=C1-AVER(C1)
8  LET C2=C2-AVER(C2)
9  LET C3-C3-AVER(3)
10 COPY C1-C3 INTO M1
11 TRAN M1, M2
12 MULT M2 M1, M3
```

```
13 PRINT M3
14 DIAG C4, M4
15 MULT M4 M3, M4
16 PRINT M4
17 STOP
```

6.5 Some MINITAB examples for multivariate analysis

6.5.1 Example of eigenanalysis of non-symmetric matrix

```
PRINT M1
   MATRIX M1                      4 ROWS BY        4 COLUMNS
```

$$\begin{bmatrix} 1.23981 & 1.11477 & 0.28177 & 0.32404 \\ 1.70016 & 1.52869 & 0.38640 & 0.44436 \\ -0.52596 & -0.47290 & -0.11954 & -0.13747 \\ 9.21021 & 8.28137 & 2.09321 & 2.40721 \end{bmatrix} = W^{-1} A$$

```
* MULT M1 BY M1,M2
* MULT M2 BY M2,M2
* MULT M2 BY M2,M2          "Powering" the matrix
* MULT M2 BY M2,M2
* MULT M2 BY M1,M3
* COPY M2 TO C11-C14
* COPY M3 TO C15-C18
* LET K3-(SUM(C15))/(SUM(C11))
* PRINT K3
   K3         5.05618        = the first root
* LET C10-C15/C11
* PRINT C10

COLUMN       C10
COUNT          4
     5.05618       5.05618      5.05618       5.05618 - check

* LET C10=C11/10000
* PRINT C10
COLUMN       C10
```

```
COUNT          4
     4473961.      6135180.   -1897976.     33235921.

*LET C10=C10/SQRT(SUM(C10*C10))
*PRINT C10
COLUMN       C10
COUNT          4
    [0.131028      0.179680   -0.055586     0.973374]

     = the vector associated with the 1st root
```

6.5.2 Example of calculation of determinant of a matrix

```
*PRINT M1

  MATRIX M1                    3 ROWS BY       3 COLUMNS

   [1.00000      0.70000     0.80000]
   [0.70000      1.00000     0.60000]
   [0.80000      0.60000     1.00000]

* EIGEN M1,C1,M2
* LET C2=LOGE(C1)
* LET K1=EXPO(SUM(C2))
* PRINT K1
   K1        0.182000      - the determinant
```

6.5.3 Test of sphericity on a correlation matrix.

```
PRINT K1-K3
   K1       0.182000       = determinant of the matrix
   K2       49.0000        = number samples less one
   K3       3.00000        = number variables
* LET K4=-(K2-(2*K3+5)/6)*LOGE(K1)
* LET K5=K3*(K3-1)/2
* PRINT  K4-K5
  K4         80.3601 = X2
  K5         3.00000 = df
```

7. ORDINATION AND CLUSTER ANALYSIS

7.1 Tutorial/assignment

The data set to be used will be 'SEDABC DATA'. These are sediment samples obtained by grabs from 10-20m depth (below mean low water) at 3 locations 1 km apart in the lower Bay of Fundy on the Atlantic coast of Canada (where these samples were taken, the tidal range is about 18m). There are 60 samples (n=60), 20 from each of the 3 locations A, B and C. There are 4 variables: % sand, % silt-clay, % gravel, and organic content as % of total dry weight. The first 3 variables add to 100%. The 5th column of data contains "location codes": 1=A, 2=B, and 3=C.

To begin with, we will ignore the fact that we know that the samples come from 3 locations. We will treat the data as "unpartitioned" for purposes of analysis, and we will apply a principal components analysis, a cluster analysis, and the "variable subset seection" FORTRAN program 'RSLCTIBM FORTRAN' (based on an algorithm originally proposed by L. Orloci). Each of these methods somehow "look for" partitions of the data in order to describe the structure in the data. After these analyses we will "remember" that the samples come from different locations and we will see whether the structure that has been described is related to the locations.

We will use MINITAB, SAS, APL, and a FORTRAN program. The Orloci & Kenkel Apple DOS 3.3 BASIC programs also include programs analagous to those we will use. We will not use them as part of this tutorial/assignment. However some of you may wish to try them if you are going to be limited to BASIC programs "back home".

7.1.1 MINITAB

Run the MINITAB example of doing a PCA (a handout you have already been given). Include the "sphericity test" insert it just after you do the "EIGEN---" command. (Choose your own C, K, and M numbers so they do not conflict with the PCA analysis!) If you have problems with the sphericity test because one of the roots is zero, then drop the 4th root which is zero and do the sphericity test using only the 3 non-zero roots. Run interactively first, then as a batch job, and then print out the 'fn MINITAB' and 'fn OUTPUT'.

7.1.2 APL

a. Now go into APL. If you have not yet done so, read the descriptions of functions MATFORM, COVAR, GEIG, and ISOTROPY (by entering each name with "DES" appended). If that is unclear, enter "DESCRIBEFNS".

b. The workspace UNESCO also contains the variable SEDABC. Enter 'SEDABC' and you will see the same data set as in the file 'SEDABC DATA'. The function MATFORM was used to enter the data and shape them into this 60-by-5 matrix.

c. Run COVAR using the option to create a covariance matrix (enter '0 COVAR SEDABC[;1 2 3 4]'). Do you understand the bracketed part? If not, just enter 'SEDABC[;1 2 3 4]' and compare the response with the response you get when you enter 'SEDABC'. Rename the covariance matrix from M to MC (enter 'MC←M').
   Run COVAR again, this time with the option to create a correlation matrix (enter '1 COVAR SEDABC[;1 2 3 4]'). Rename it to MR (enter 'MR←M').

d. Now run GEIG on the covariance matrix by entering 'GEIG MC'. Write down this root (which is for PC I) and its associated vector, then follow the instructions and continue by entering 'GEIG N'. Write down the root and vector for PC II. Again enter 'GEIG N' and write down the root and vector for PC III. Sum the 3 roots. Enter 'MC' and sum the diagonal elements (the variances) in the covariance matrix. Are they the same? If they are, then the 4th root is zero (as you know it is from the MINITAB analysis), so there is no point in doing 'GEIG N' again.

e. Repeat (d) on the correlation matrix (stored in MR). Also do the sphericity test on the correlation matrix, by entering '60 ISOTROPY roots', where 60 is the number of samples from which the correlation matrix was calculated and "roots" is a vector containing the roots. Again, you can not include a zero root so leave out the 4th root, which is zero.

f. Compare all your APL results with those you obtained using MINITAB.

7.1.3 SAS

a. Now prepare a file named 'SEDABC SAS', using XEDIT. Your file should look like this:

```
TITLE SAS ANALYSIS ON SEDIMENT DATA;
DATA SEDABC;
INPUT PERSAND PERSLTCL PERGRAV PERORG LOCATION;
CARDS;
```

(the SEDABC data go here - use the 'GET SEDABC DATA' command)

```
PROC PRINT;
PROC PLOT; PLOT PERSAND*PERSLTCL=LOCATION;
PROC PLOT; PLOT PERSAND*PERGRAV=LOCATION;
PROC PLOT; PLOT PERSAND*PERORG=LOCATION;
PROC PLOT; PLOT PERSLTCL*PERGRAV=LOCATION;
PROC PLOT; PLOT PERSLTCL*PERORG=LOCATION;
PROC PLOT; PLOT PERGRAV*PERORG=LOCATION;
PROC PRINCOMP OUT=COVPCS COV; VAR PERSAND PERSLTCL PERGRAV
PERORG;
PROC PRINCOMP DATA=SEDABC OUT=CORPCS; VAR PERSAND PERSLTCL
PERGRAV PERORG;
PROC PLOT DATA=COVPCS; PLOT PRIN1*PRIN2=LOCATION;
PROC PLOT DATA=CORPCS; PLOT PRIN1*PRIN2=LOCATION;
PROC CLUSTER DATA=SEDABC OUTTREE=TREE;
VAR PERSAND PERSLTCL PERGRAV PERORG; ID LOCATION;
PROC PRINT DATA=TREE;
PROC PLOT; PLOT _CCC_ * _NCL_;
```

b. Run the SAS job, and look at the output (enter 'TY SEDABC LISTING'). Compare the correlations between the variables with the bivariate plots of the variables. Compare the bivariate plots of the variables with the "PC I vs. PC II" plots. How do the PCA and cluster analysis results relate to the three locations?

7.1.4 FORTRAN

Now we will run the FORTRAN program 'RSLCTIBM FORTRAN'. This is the algorithm which selects a subset of variables, such that the subset best represents (is most highly correlated with) the whole set.

a. Enter 'TY RSLCTIBM DATA'and observe that the data are the same as 'SEDABC DATA'.

b. Enter 'TY RSLCTIBM FORTRAN' and read the comment lines that proceed the program itself.

c. Now run the program, by entering 'FORTVS RSLCTIBM'. Wait for the final "R;--------", and then enter 'TY RSLCTIBM OUTPUT'. Read it and try to understand it.

d. Note the first 2 variables "selected". What percent of the total correlation structure do they account for? What percent of the correlation structure did the first two principal components (PCs) account for (refer to MINITAB, SAS or APL runs)? Are the 2 "best variables" almost as good at accounting for correlation structure as the 2 best linear combinations of all 4 variables (that is one way of saying what PCs are)? Look at the SAS bivariate plots. Does the bivariate plot of the 2 best variables against each other show the most information? Does it show low or high correlation?

e. Look at the vectors associated with the first 2 PCs. Is the coefficient associated with the "best variable" in the PC I vector relatively large in magnitude? Is the coefficient associated with the "2nd best variable" in the PC II vector relatively large in magnitude?

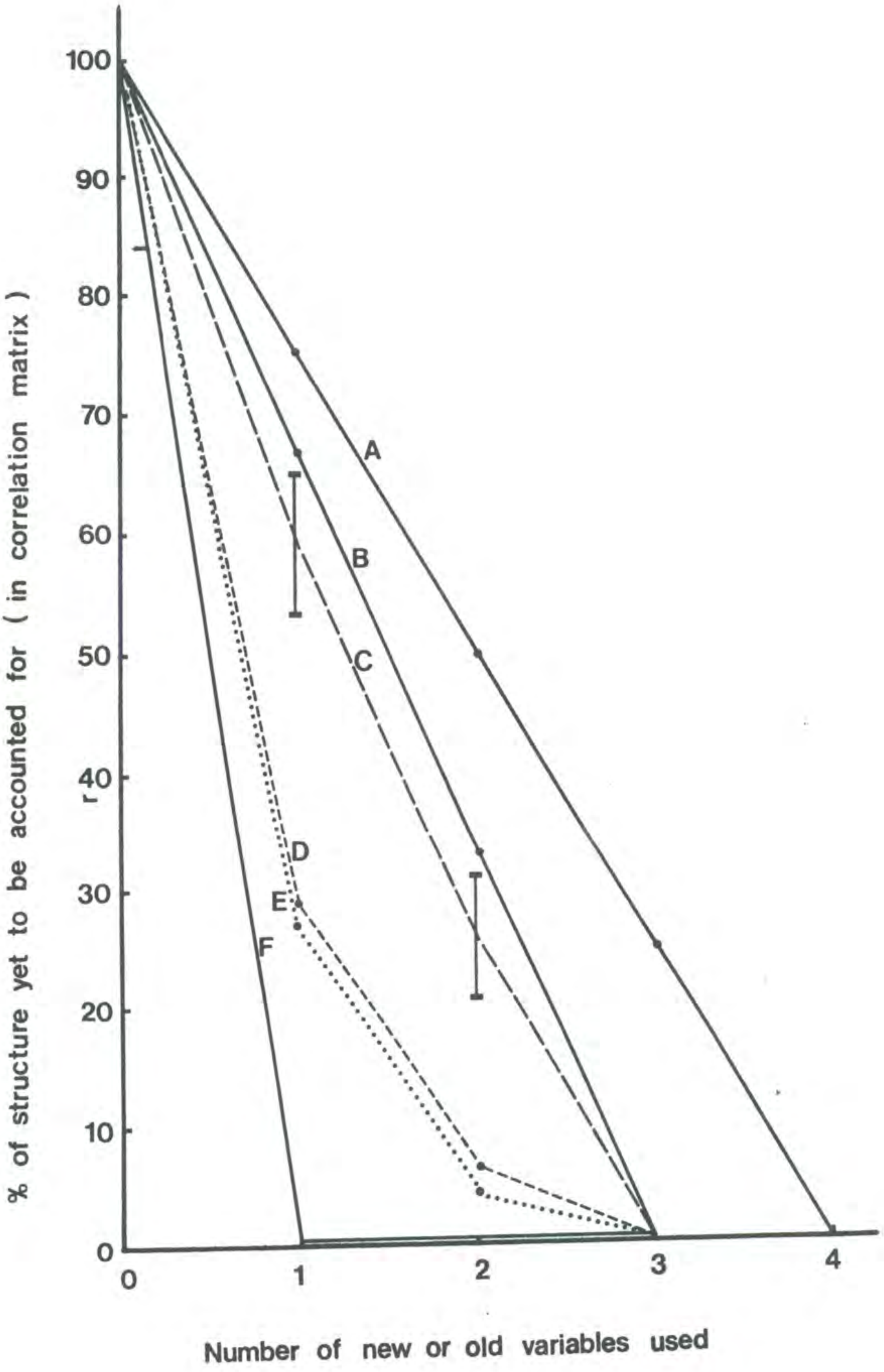
7.1.5 Overall evaluation.

Now try to evaluate all this. You used 4 analytical approaches to evaluate the correlation structure in this n-60-by-p-4 data matrix, and you ignored the information contained in a 5th "location code" variable. Try to answer the following questions:

a. Does "location" appear to be reacted to, or involved in, the correlation structure you described when ignoring the location information? Try to interprete any relationship you see.

b. You used 4 analytical methods or approaches: (1) the sphericity test of the Ho: "no nonzero correlations" which is equivalent to Ho:"no correlation structure"; (2) principal components analysis (PCA) which finds new variables (new axes) which most efficiently display the structure in the data;(3) RSLCT which selects the best of the original variables for display of the structure in the data; and (4) cluster analysis which finds the best groups of samples to describe the structure. Can you see how they are describing (testing in the case of the sphericity test) the same structure in this data set, though in different ways? Which do you think does the best job (go ahead and be subjective!)?

c. You used MINITAB, SAS, APL, and FORTRAN program. (You may also have used some of the Orloci & Kenkel Apple DOS 3.3 programs.) Can you see that the results (e.g. PCA, sphericity test) are basically the same when done by the different languages or packages? Do you have likes and dislikes related to ease of use, clarity of output, or any other characteristic?


% of structure yet to be accounted for (in correlation matrix)
100
90
80
70
60
50
40
30
20
10
0
0
1
2
3
4
Number of new or old variables used
A
B
C
D
E
F

A : No redundancy in 4 dimensions
= all 4 roots exactly equal
= sphericity in 4 dimensions
(all zero correlations)

B : No redundancy in 3 dimensions
= 1 root zero, rest exactly equal
= sphericity in 3 dimensions
(all zero correlations)

C : Expected when random
uncorrelated data are
used (nonzero correlations
by chance only).
.95 cls on a single run
are shown.
(In 3 dimensions.)

D : RSLCT on sediment data.

E : PCA on sediment data.

F : Total redundancy
= only 1 nonzero root
= all perfect correlations.

7.2. Job Listings and Outputs.

```
FILE: PCA2     MINITAB  A1  VM/SP - CONVERSATIONAL MONITOR SYSTEM                    PAGE 001


READ C1-C5
     2.350   97.150    0.0       9.325 1.
     4.522   94.250    1.123     3.479 1.
     3.564   95.944    0.472     8.940 1.
     3.117   96.592    0.191     8.753 1.
     1.513   94.170    4.317     8.451 1.
     2.520   97.480    0.0       6.377 1.
     2.394   96.120    0.296     7.058 1.
     2.460   97.467    0.073     5.874 1.
    31.548   60.702    7.750     5.190 1.
     5.695   93.341    0.454     7.330 1.
     2.365   95.763    0.172     3.734 1.
     8.561   90.098    1.341     5.192 1.
     4.237   95.593    0.170     7.303 1.
     4.329   95.671    0.0       7.560 1.
     2.098   97.344    0.553     7.034 1.
     1.761   98.123    0.166     7.397 1.
     2.309   97.691    0.0       8.377 1.
     4.385   94.620    0.994     7.530 1.
     3.295   95.477    1.228     5.972 1.
     2.953   97.047    0.0       8.691 1.
    46.607   53.003    0.390     5.238 2.
    74.874   24.588    0.538     2.356 2.
    80.096   19.211    0.693     2.412 2.
    81.449   17.401    1.150     2.236 2.
    78.150   20.690    1.160     2.589 2.
    49.475   49.396    1.129     5.073 2.
    47.414   52.418    0.168     5.709 2.
    86.558   12.607    0.835     4.360 2.
    50.699   48.619    0.682     2.275 2.
    80.388   18.146    1.465     2.259 2.
    58.262   41.068    0.670     3.572 2.
    60.382   38.412    1.206     3.933 2.
    55.870   43.314    0.816     2.243 2.
    74.645   23.567    1.788     2.639 2.
    74.699   23.581    1.720     2.705 2.
    42.257   57.207    0.536     5.542 2.
    41.967   57.484    0.549     6.299 2.
    72.371   27.043    0.586     2.055 2.
    76.056   23.602    0.342     2.563 2.
    80.534   18.572    0.894     2.534 2.
     0.650   99.350    0.0       7.909 3.
     0.600   99.400    0.0       7.364 3.
     0.550   99.450    0.0       6.337 3.
     0.985   99.015    0.0       6.375 3.
     0.704   99.296    0.0       7.228 3.
     1.208   98.792    0.0       4.300 3.
     1.469   98.531    0.0       7.524 3.
     0.855   99.145    0.0       7.914 3.
     2.193   97.807    0.0       7.535 3.
     1.832   98.168    0.0       7.298 3.
     0.960   99.005    0.035     6.396 3.
     0.754   99.193    0.053     6.333 3.
     1.150   98.850    0.0      11.752 3.
     2.078   97.922    0.0       7.526 3.
```

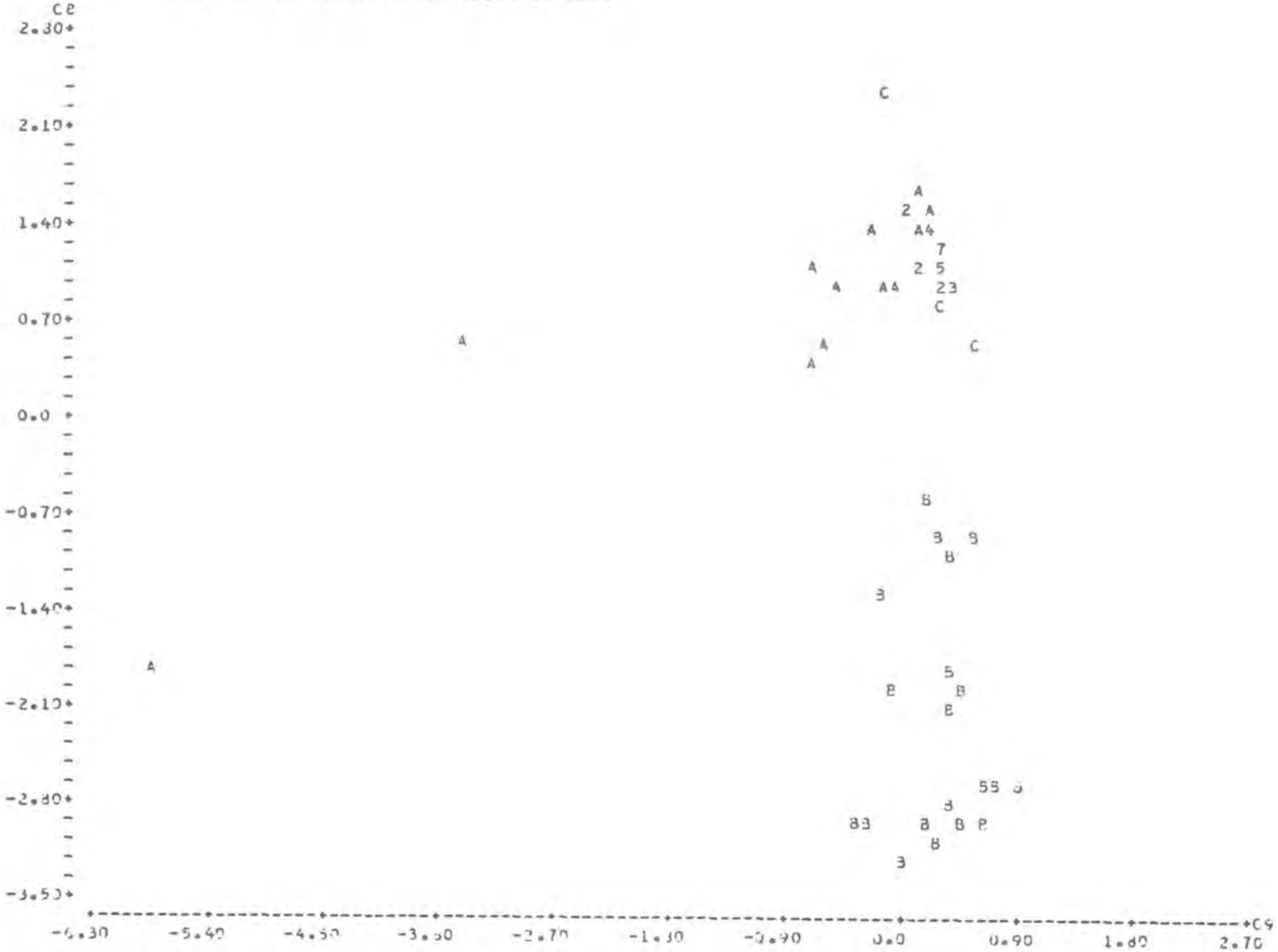
```
   1.053  93.347   0.0      7.490 3.
   0.344  99.126   0.030    6.069 3.
   1.669  93.332   0.0      6.096 3.
   1.501  93.179   0.021    5.997 3.
   1.087  93.710   0.0      6.629 3.
   2.504  97.391   0.195    5.354 3.
PRINT C1-C5
CORR C1-C4, M1
PRINT M1
NOTE M1 IS THE CORRELATION MATRIX
EIGEN M1, C6, M2
PICK 1 3 C6, C7
NOTE C7 CONTAINS THE FIRST THREE (NON-ZERO) ROOTS OF THE CORR MATRIX
LET C8=LOGE(C7)
NOTE C8 CONTAINS THE LOG OF THE FIRST 3 ROOTS OF THE CORRELATION MATRIX
LET K1=EXPO(SUM(C8))
PRINT K1
NOTE K1 IS THE PRODUCT OF THE FIRST THREE ROOTS OF THE CORRELATION MATRIX
LET K2=59
NOTE K2 IS THE NUMBER OF OBSERVATIONS MINUS ONE
LET K3=4
NOTE K3 IS THE NUMBER OF VARIABLES
LET K4=-(K2-(2*K3+5)/6)*LOGE(K1)
NOTE K4 IS THE CHI-SQUARE VALUE FOR SPHERICITY TEST
LET K5=K3*(K3-1)/2
NOTE K5 IS THE DEGREES OF FREEDOM
PRINT K4-K5
PRINT C6
NOTE C6 GIVES THE EIGENVALUES OR ROOTS OF THE CORRELATION MATRIX
SUM C6, K1
NOTE K1 IS THE SUM OF THE EIGENVALUES, AND SHOULD HAVE VALUE 4
LET C7=100*C6/4
PRINT C7
NOTE C7 ARE THE EIGENVALUES GIVEN IN PERCENTAGE
PRINT M2
NOTE M2 GIVES THE EIGENVECTORS OF THE CORRELATION MATRIX
LET C1=(C1-AVER(C1))/STAN(C1)
LET C2=(C2-AVER(C2))/STAN(C2)
LET C3=(C3-AVER(C3))/STAN(C3)
LET C4=(C4-AVER(C4))/STAN(C4)
NOTE C1-C4 NOW CONTAIN THE Z TRANSFORMATION OF THE ORIGINAL DATA
COPY C1-C4 INTO M3
PRINT M3
NOTE M3 IS THE MATRIX CONTAINING THE STANDARDISED VALUES OF THE DATA
MULT M3 M2, M4
COPY M4 INTO C8-C11
DESCRIBE C8-C11
NOTE C8-C11 ARE THE PCI-PCIV
WIDTH 100, 50
LPLOT C8 C9, C5
NOTE THIS GIVES THE PLOT OF PCI VERSUS PCII
STOP
```

```
FILE: SEDIMENT OUTPUT   A1  VM/SP - CONVERSATIONAL MONITOR SYSTEM                                          PAGE 001

1MINITAB RELEASE 81.1 *** COPYRIGHT - PENN STATE UNIV. 1981
 MAY  4, 1985 *** NATIONAL UNIVERSITY OF SINGAPORE - LOCAL VERSION 02/12/1982
 STORAGE AVAILABLE    4300

  PRINCIPAL COMPONENTS ANALYSIS ON SEDIMENT PARAMETERS
  OF CANADIAN GRAB SAMPLES USING MINITAB
 --
 COLUMN        C1           C2           C3           C4            C5
 COUNT         60           60           60           60            60
 ROW        % SAND     % SILT-CLAY   % GRAVEL     % O.M.        AREA CODE
   1        2.8500      97.1500     0.0          9.3250           1.
   2        4.6220      94.2500     1.12800      8.4790           1.
   3        3.5640      95.9440     0.49200      8.8400           1.
   4        3.1170      96.6920     0.19100      9.7580           1.
   5        1.5130      94.1700     4.31700      8.4510           1.
   6        2.5200      97.4300     0.0          6.8790           1.
   7        2.8940      96.8200     0.28600      7.0580           1.
   8        2.4600      97.4670     0.07300      6.8940           1.
   9       31.5480      60.7020     7.75000      6.1900           1.
  10        5.6950      93.8410     0.46400      7.3300           1.
  11        2.8650      96.9630     0.17200      8.7340           1.
  12        8.5610      90.0980     1.34100      6.1920           1.
  13        4.2370      95.5930     0.17000      7.3030           1.
  14        4.3290      95.6710     0.0          7.5600           1.
  15        2.0980      97.3440     0.55800      7.0340           1.
  16        1.7610      98.1230     0.11600      7.8970           1.
  17        2.3090      97.6910     0.0          8.3770           1.
  18        4.3360      94.6200     0.97400      7.5300           1.
  19        3.2950      95.4770     1.22800      5.9720           1.
  20        2.9530      97.0470     0.0          8.6910           1.
  21       46.6070      53.0030     0.39000      5.2380           2.
  22       74.8740      24.5880     0.53800      2.8560           2.
  23       80.0960      19.2110     0.69300      2.4120           2.
  24       81.4490      17.4010     1.15000      2.2360           2.
  25       78.1500      20.6900     1.16000      2.6890           2.
  26       49.4750      49.3960     1.12900      5.0730           2.
  27       47.4140      52.4130     0.16890      5.7090           2.
  28       86.5580      12.6070     0.83500      4.3600           2.
  29       50.6990      48.6190     0.68200      2.2750           2.
  30       85.3880      13.1460     1.46600      2.2590           2.
  31       58.2620      41.0680     0.67000      3.5920           2.
  32       59.3820      39.4120     1.20600      3.9330           2.
  33       55.8700      43.3140     0.81600      2.2430           2.
  34       74.6450      23.5670     1.78800      2.6380           2.
  35       74.6990      23.5810     1.72000      2.7050           2.
  36       42.2570      57.2070     0.53600      5.5420           2.
  37       41.9570      57.4940     0.54900      6.2990           2.
  38       72.3710      27.0430     0.58600      2.0550           2.
  39       76.0560      23.6020     0.34200      2.6530           2.
  40       80.5340      18.5720     0.89400      2.5840           2.
  41        0.6500      99.3500     0.0          7.9080           3.
  42        0.6000      99.4000     0.0          7.9640           3.
  43        0.5500      99.4500     0.0          6.3370           3.
  44        0.9850      99.0150     0.0          6.8250           3.
  45        0.7040      99.2960     0.0          7.2280           3.
```

```
FILE: SEDIMENT OUTPUT   A1  VM/SP - CONVERSATIONAL MONITOR SYSTEM

            % SAND    % SILT-CLAY   % GRAVEL       % O.M.       AREA CODE
  46        1.2080     98.7920      0.0           4.3000           3.
  47        1.4690     98.5310      0.0           7.5240           3.
  48        0.8550     99.1450      0.0           7.9140           3.
  49        2.1930     97.8070      0.0           7.5350           3.
  50        1.8320     98.1680      0.0           7.2980           3.
  51        0.9600     99.0050      0.03500       6.3960           3.
  52        0.7540     99.1880      0.05800       6.3330           3.
  53        1.1500     98.8500      0.0          11.7620           3.
  54        2.0780     97.9220      0.0           7.5260           3.
  55        1.6530     98.3470      0.0           7.4900           3.
  56        0.8440     99.1260      0.03000       6.0680           3.
  57        1.6680     98.3320      0.0           6.0860           3.
  58        1.8010     98.1780      0.02100       6.9970           3.
  59        1.0870     98.9130      0.0           6.6290           3.
  60        2.5040     97.3910      0.10500       5.8540           3.

--
  THE FOLLOWING ARE THE CORRELATION COEFFICIENTS BETWEEN
  CI AND CJ

              C1        C2        C3
C2          -0.999
C3           0.275   -0.309
C4          -0.861    0.859   -0.194

--
   M1 IS THE CORRELATION MATRIX OF THE SEDIMENT PARAMETERS

   MATRIX M1                  4 ROWS BY      4 COLUMNS

     1.00000     -0.99936      0.27496     -0.86114
    -0.99936      1.00000     -0.30918      0.85901
     0.27496     -0.30918      1.00000     -0.19412
    -0.86114      0.85901     -0.19412      1.00000
--
   K1          0.469345   DETERMINANT OF THE 3-ROOT DIAGONAL MATRIX
--
   K4           42.7897   CHI-SQUARE VALUE FOR SPHERICITY TEST
   K5           6.00000   ASSOCIATED DEGREES OF FREEDOM
--
  C6 GIVES THE EIGENVALUES OR ROOTS OF THE CORRELATION MATRIX
--
COLUMN         C6
COUNT           4
     2.92073      0.90039      0.17887      0.00000

--
   NOTE THAT THE SUM OF THE ROOTS EQUALS THE SUM OF THE
   DIAGONAL ELEMENTS OF THE CORRELATION MATRIX, I.E.
   SUM      =        4.0000
--


--
```

```
 C7 GIVES THE ROOTS OR EIGENVALUES IN PERCENTAGE
COLUMN        C7
COUNT          4
     73.0133     22.5223      4.4593      0.0000

--
  M2 IS THE MATRIX OF EIGENVECTORS OF THE CORRELATION MATRIX
   MATRIX M2                   4 ROWS BY     4 COLUMNS

   -0.572881    0.116192    0.405130    0.702979
    0.575171   -0.079355   -0.397287    0.710719
   -0.223838   -0.968759   -0.091743    0.026452
    0.537211   -0.204421    0.813301   -0.000014

--
--
  M3 GIVES THE STANDARDISED VALUES OF THE ORIGINAL DATA
   MATRIX M3                  60 ROWS BY     4 COLUMNS

    -0.67544     0.68756    -0.52371     1.43570
    -0.61867     0.59555     0.43693     1.06141
    -0.55256     0.64934    -0.10471     1.22112
    -0.66689     0.67304    -0.36105     1.18484
    -0.71223     0.59312     3.15279     1.04902
    -0.53601     0.69301    -0.52371     0.35352
    -0.67403     0.67710    -0.29014     0.43272
    -0.68794     0.69760    -0.46154     0.36016
     0.24404    -0.45752     6.07645     0.04869
    -0.58429     0.58269    -0.12855     0.55306
    -0.67496     0.68163    -0.37723     1.17423
    -0.49246     0.46407     0.61833     0.04957
    -0.63100     0.63821    -0.27893     0.54111
    -0.62805     0.64067    -0.52371     0.65492
    -0.69953     0.69373    -0.04850     0.42210
    -0.71033     0.71839    -0.33234     0.80391
    -0.69277     0.70470    -0.52371     1.01628
    -0.62623     0.60739     0.32281     0.64154
    -0.66119     0.63454     0.52210    -0.04776
    -0.67214     0.68429    -0.52371     1.15520
     0.72653    -0.71131    -0.17157    -0.37250
     1.63221    -1.61201    -0.03553    -1.42636
     1.79952    -1.79241     0.05647    -1.62230
     1.84237    -1.83977     0.45567    -1.70067
     1.73717    -1.73554     0.46419    -1.50025
     0.31842    -0.32532     0.43779    -0.44550
     0.75239    -0.73025    -0.38063    -0.16412
     2.00656    -1.99170     0.18740    -0.76095
     0.85764    -0.85044     0.05710    -1.68341
     1.80883    -1.81516     0.72479    -1.67049
     1.09996    -1.08974     0.04689    -1.10074
     1.16748    -1.17391     0.57336    -0.94937
     1.02332    -1.01836     0.17122    -1.69757
     1.62457    -1.64436     0.97931    -1.52281
     1.62660    -1.64392     0.94110    -1.49317
     0.53716    -0.57329    -0.06723    -0.23900
     0.57787     -0.56950     -0.05616      0.09691
     1.55201     -1.53421     -0.02465     -1.78075
     1.67008     -1.64325     -0.23245     -1.51175
     1.81355     -1.80256      0.23765     -1.54670
    -0.74593      0.75729     -0.52371      0.80878
    -0.74753      0.75986     -0.52371      0.78931
    -0.74913      0.76045     -0.52371      0.11373
    -0.73519      0.74666     -0.52371      0.36060
    -0.74420      0.75557     -0.52371      0.50793
    -0.72805      0.73959     -0.52371     -0.78750
    -0.71969      0.73132     -0.52371      0.63899
    -0.73936      0.75078     -0.52371      0.81144
    -0.69649      0.70338     -0.52371      0.64376
    -0.70806      0.71962     -0.52371      0.53990
    -0.73600      0.74534     -0.49390      0.13983
    -0.74260      0.75214     -0.47431      0.11196
    -0.72991      0.74143     -0.52371      2.51390
    -0.70018      0.71202     -0.52371      0.63977
    -0.71379      0.72549     -0.52371      0.62385
    -0.73971      0.75018     -0.47915     -0.00529
    -0.71331      0.72562     -0.52371      0.00268
    -0.70905      0.72014     -0.50532      0.40573
    -0.73193      0.74343     -0.52371      0.24292
    -0.68653      0.69519     -0.43429     -0.09997

--
 C8-C11 ARE THE PRINCIPAL COMPONENTS, I.E. PCI-PCIV
 THE DESCRIPTIVE STATISTICS FOR THE FOUR PRINCIPAL COMPONENTS ARE:

   C8           N =   60     MEAN = 0.000011224      ST.DEV. =        1.71
   C9           N =   60     MEAN =-0.000005957      ST.DEV. =       0.949
   C10          N =   60     MEAN = 0.000010253      ST.DEV. =       0.422
   C11          N =   60     MEAN = 0.000007924      ST.DEV. =  0.000144
--
```

PLOT OF PCI VERSUS PCII, CODED BY AREA



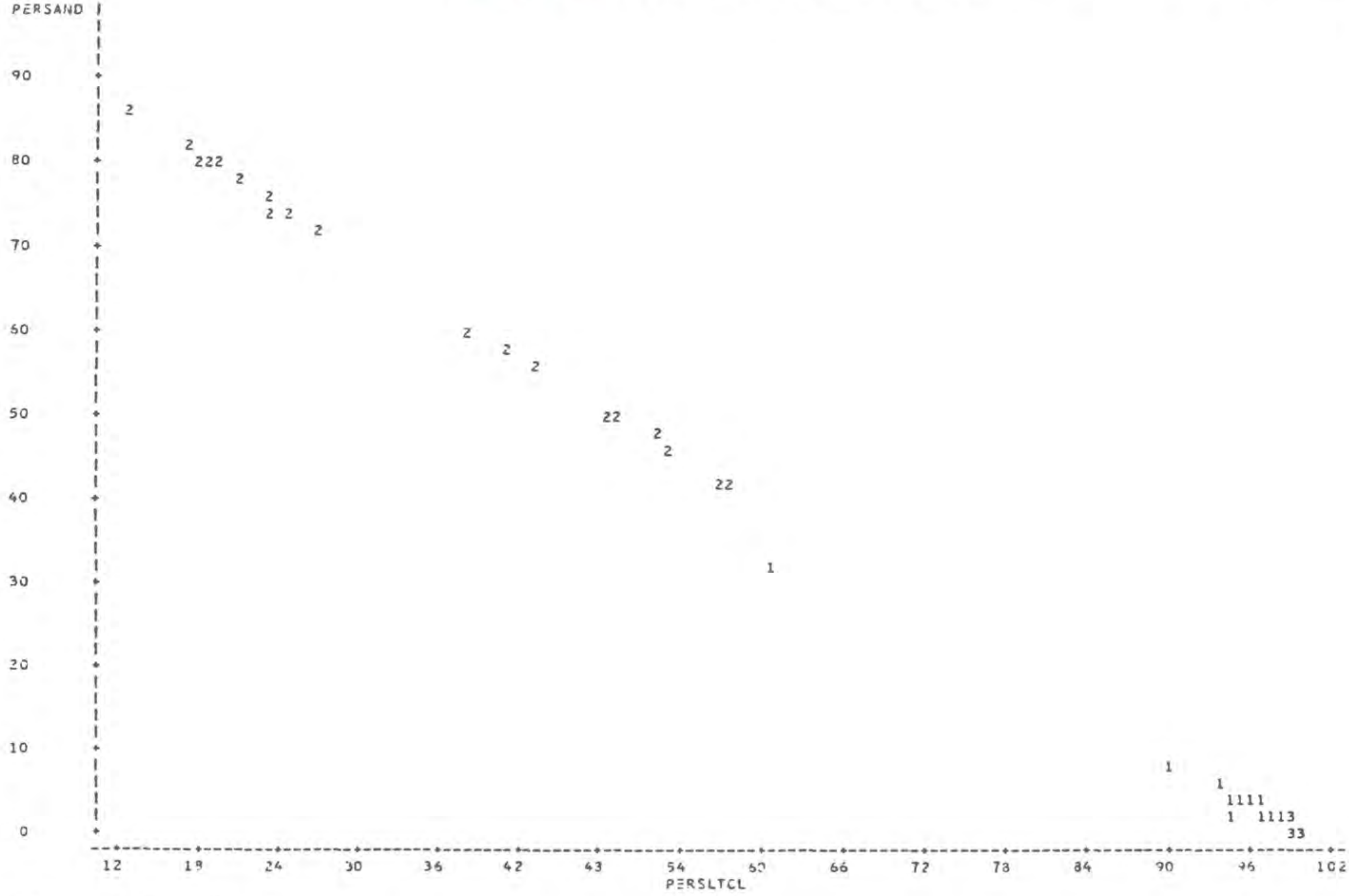
```
*** MINITAB *** STATISTICS DEPT * PENN STATE UNIV. * RELEASE 81.1 *
STORAGE AVAILABLE     4300
```

```
TITLE SAS ANALYSIS ON SEDIMENT DATA;
DATA SEDABC;
INPUT PERSAND PERSLTCL PERGRAV PERORG LOCATION;
CARDS;
    2.850  97.150   0.0     9.325 1.
    4.622  94.250   1.128   3.477 1.
    3.564  95.944   0.492   8.940 1.
    3.117  96.692   0.191   3.753 1.
    1.513  94.170   4.317   3.451 1.
    2.520  97.480   0.0     6.379 1.
    2.374  96.320   0.296   7.053 1.
    2.460  97.467   0.073   6.394 1.
   31.548  60.702   7.750   6.190 1.
    5.695  93.341   0.464   7.330 1.
    2.865  96.963   0.172   9.734 1.
    3.561  95.098   1.341   6.192 1.
    4.237  95.593   0.170   7.303 1.
    4.329  95.671   0.0     7.560 1.
    2.093  97.344   0.563   7.034 1.
    1.751  98.123   0.156   7.337 1.
    2.309  97.691   0.0     8.377 1.
    4.386  94.520   0.974   7.530 1.
    3.295  95.477   1.228   5.972 1.
    2.953  97.047   0.0     8.691 1.
   46.607  53.003   0.390   5.233 2.
   74.874  24.588   0.538   2.356 2.
   80.096  19.211   0.693   2.412 2.
   81.449  17.401   1.150   2.236 2.
   78.150  20.690   1.160   2.689 2.
   49.475  49.396   1.129   5.073 2.
   47.414  52.413   0.153   5.709 2.
   86.558  12.607   0.835   4.360 2.
   50.699  43.519   0.632   2.275 2.
   80.388  18.146   1.466   2.259 2.
   58.262  41.068   0.670   3.592 2.
   60.382  38.412   1.206   3.733 2.
   55.870  43.314   0.816   2.243 2.
   74.645  23.567   1.788   2.633 2.
   74.699  23.581   1.720   2.705 2.
   42.257  57.207   0.536   5.542 2.
   41.967  57.484   0.549   6.299 2.
   72.371  27.043   0.586   2.055 2.
   76.056  23.602   0.342   2.663 2.
   80.534  18.572   0.894   2.584 2.
    0.650  99.350   0.0     7.908 3.
    0.600  99.400   0.0     7.864 3.
    0.550  99.450   0.0     6.337 3.
    0.985  99.015   0.0     6.895 3.
    0.704  99.296   0.0     7.220 3.
    1.208  98.792   0.0     4.300 3.
    1.469  98.531   0.0     7.524 3.
    0.855  99.145   0.0     7.914 3.
    2.193  97.807   0.0     7.535 3.
    1.832  98.168   0.0     7.278 3.
    0.960  99.005   0.035   6.396 3.
   0.754  99.183   0.053    5.333 3.
   1.150  98.850   0.0     11.762 3.
   2.078  97.922   0.0      7.526 3.
   1.653  98.347   0.0      7.490 3.
   0.844  99.123   0.033    5.063 3.
   1.668  98.332   0.0      5.095 3.
   1.801  98.178   0.021    6.997 3.
   1.087  98.913   0.0      6.629 3.
   2.504  97.391   0.105    5.354 3.
PROC PRINT;
PROC PLOT; PLOT PERSAND*PERSLTCL=LOCATION;
PROC PLOT; PLOT PERSAND*PERGRAV=LOCATION;
PROC PLOT; PLOT PERSAND*PERORG=LOCATION;
PROC PLOT; PLOT PERSLTCL*PERGRAV=LOCATION;
PROC PLOT; PLOT PERSLTCL*PERORG=LOCATION;
PROC PLOT; PLOT PERGRAV*PERORG=LOCATION;
PROC PRINCOMP OUT=COVPCS COV; VAR PERSAND PERSLTCL PERGRAV PERORG;
PROC PRINCOMP DATA=SEDABC OUT=CORPCS; VAR PERSAND PERSLTCL PERGRAV
  PERORG;
PROC PLOT DATA=COVPCS; PLOT PRIN1*PRIN2=LOCATION;
PROC PLOT DATA=CORPCS; PLOT PRIN1*PRIN2=LOCATION;
PROC CLUSTER DATA=SEDABC OUTTREE=TREE;
     VAR PERSAND PERSLTCL PERGRAV PERORG; ID LOCATION;
PROC PRINT DATA=TREE;
PROC PLOT; PLOT _CCC_*_NCL_;
```

```
                         SAS ANALYSIS ON SEDIMENT DATA                    17:09 TUESDAY, MAY 7, 1985    1
 OBS    PERSAND    PERSLTCL    PERGRAV    PERORG    LOCATION

   1     2.850     97.150      0.000      9.325        1
   2     4.622     94.250      1.129      8.479        1
   3     3.564     95.944      0.492      8.840        1
   4     3.117     96.692      0.191      8.758        1
   5     1.513     94.170      4.317      8.451        1
   6     2.520     97.430      0.000      6.879        1
   7     2.794     96.920      0.286      7.058        1
   8     2.460     97.467      0.073      6.394        1
   9    31.548     60.702      7.750      6.190        1
  10     5.695     93.841      0.464      7.330        1
  11     2.865     96.963      0.172      8.734        1
  12     8.561     90.098      1.341      6.192        1
  13     4.237     95.593      0.170      7.303        1
  14     4.329     95.671      0.000      7.560        1
  15     2.098     97.344      0.558      7.034        1
  16     1.761     98.123      0.165      7.897        1
  17     2.309     97.691      0.000      8.377        1
  18     4.386     94.620      0.994      7.530        1
  19     3.295     95.477      1.228      5.972        1
  20     2.953     97.047      0.000      8.691        1
  21    46.607     53.003      0.390      5.239        2
  22    74.874     24.588      0.538      2.856        2
  23    80.096     19.211      0.693      2.412        2
  24    81.449     17.401      1.150      2.236        2
  25    78.150     20.690      1.160      2.689        2
  26    49.475     49.396      1.129      5.073        2
  27    47.414     52.418      0.168      5.709        2
  28    86.558     12.607      0.835      4.360        2
  29    50.699     48.619      0.682      2.275        2
  30    80.388     18.146      1.466      2.259        2
  31    58.262     41.068      0.670      3.592        2
  32    60.392     38.412      1.206      3.933        2
  33    56.370     43.314      0.316      2.243        2
  34    74.645     23.567      1.788      2.638        2
  35    74.699     23.581      1.720      2.705        2
  36    42.257     57.207      0.536      5.542        2
  37    41.957     57.484      0.559      6.239        2
  38    72.371     27.043      0.586      2.055        2
  39    76.056     23.602      0.342      2.653        2
  40    80.534     18.572      0.894      2.584        2
  41     0.650     99.350      0.000      7.908        3
  42     0.600     99.400      0.000      7.254        3
  43     0.550     99.450      0.000      6.337        3
  44     0.985     99.015      0.000      6.895        3
  45     0.704     99.296      0.000      7.228        3
  46     1.208     98.792      0.000      4.300        3
  47     1.469     98.531      0.000      7.524        3
  48     0.855     99.145      0.000      7.914        3
  49     2.193     97.807      0.000      7.535        3
  50     1.832     98.168      0.000      7.293        3
  51     0.960     99.005      0.035      6.396        3
  52     0.754     99.188      0.058      6.333        3
  53     1.150     98.850      0.000     11.752        3
  54     2.078     97.922      0.000      7.526        3
  55     1.653     98.347      0.000      7.490        3
  56     0.844     99.126      0.030      6.059        3
```

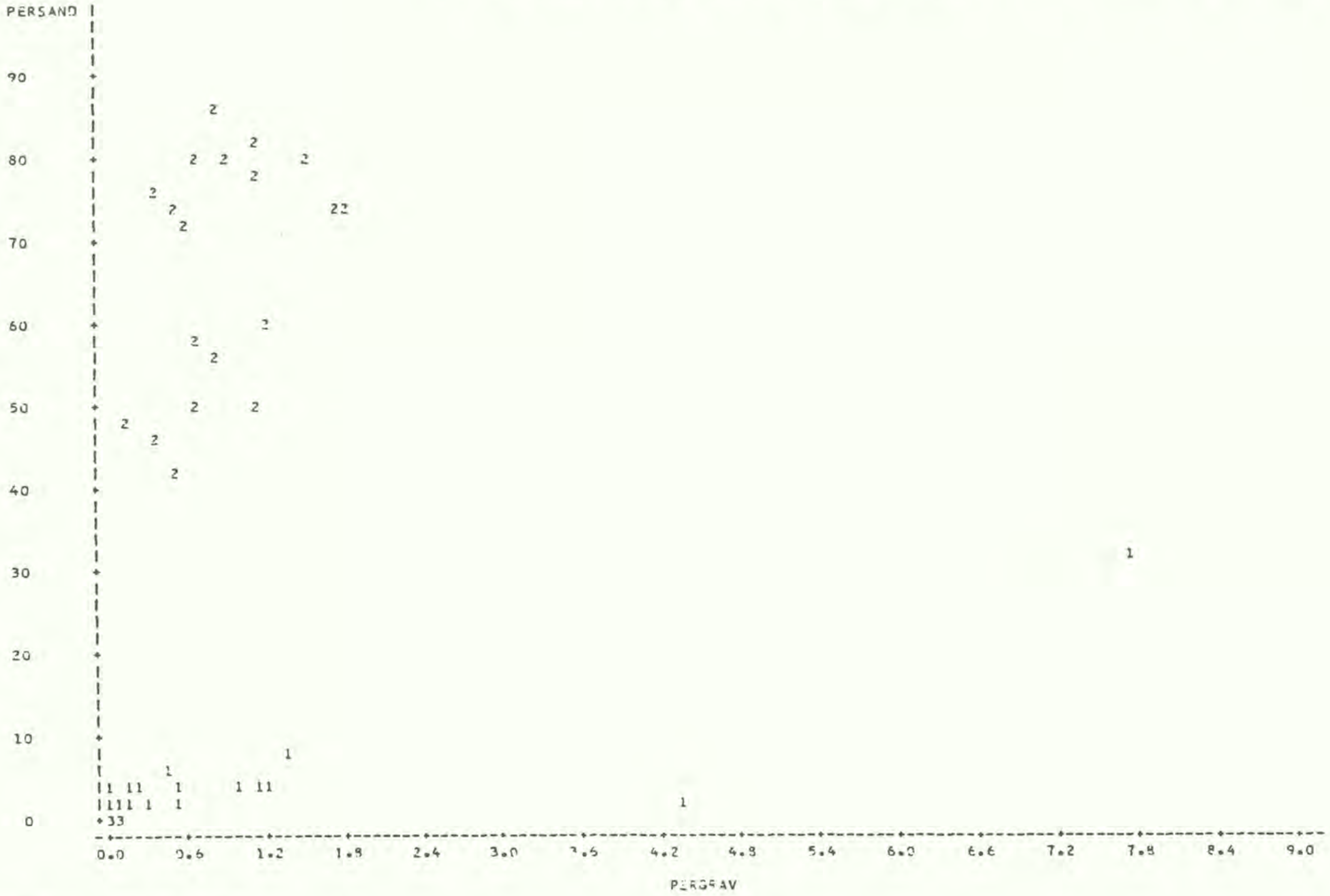
```
                    SAS ANALYSIS ON SEDIMENT DATA                            17:09 TUESDAY, MAY 7, 1985    2
OBS     PERSAND     PERSLTCL     PERGRAV     PERORG     LOCATION

57       1.668       98.332       0.000       6.086        3
58       1.801       98.178       0.021       6.997        3
59       1.087       98.913       0.000       6.629        3
60       2.504       97.391       0.105       5.354        3
```

```
                         SAS ANALYSIS ON SEDIMENT DATA                  17:09 TUESDAY, MAY 7, 1985     3
                  PLOT OF PERSAND*PERSLTCL     SYMBOL IS VALUE OF LOCATION
```



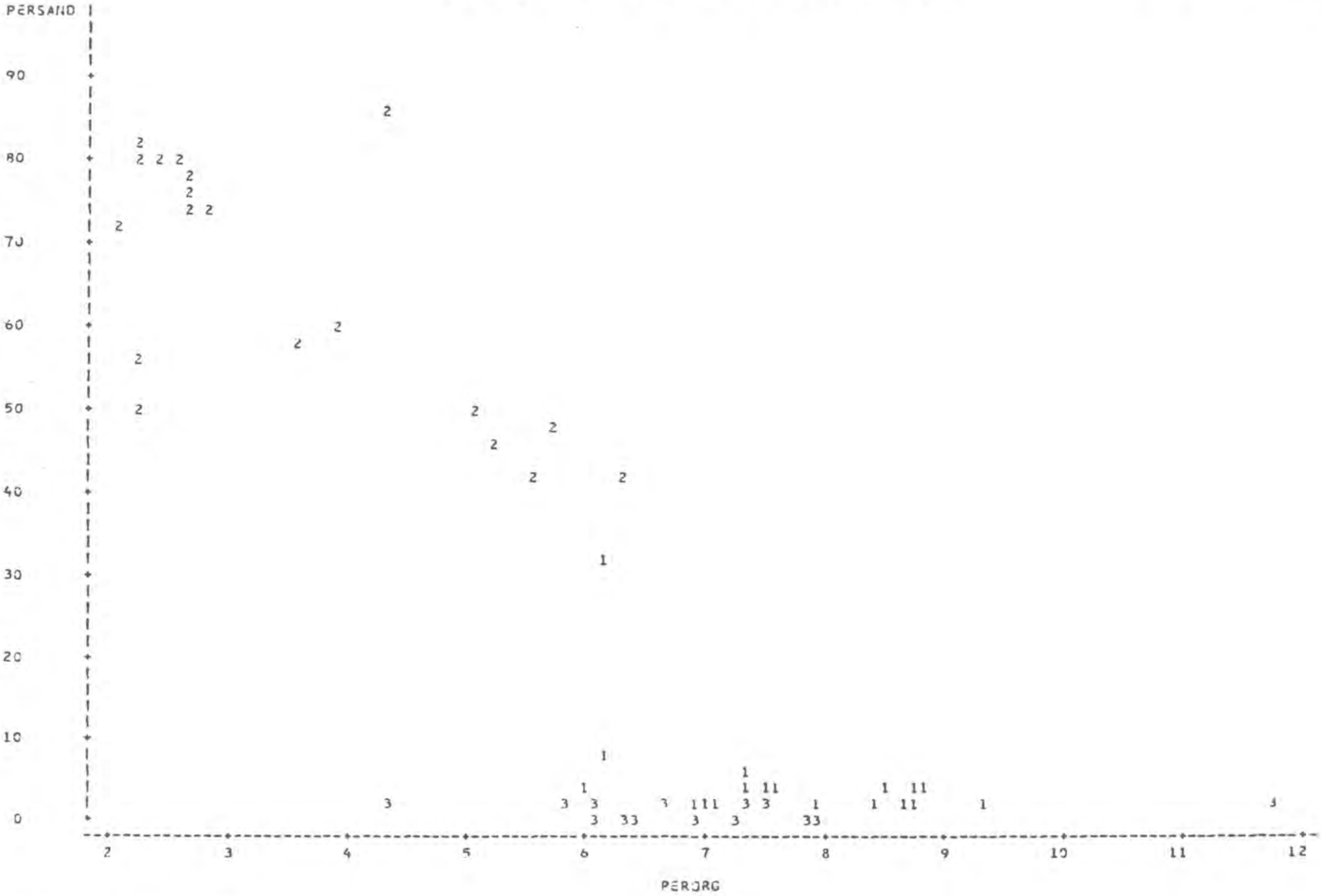
```
NOTE:      27 OBS HIDDEN
```

```
                    SAS ANALYSIS ON SEDIMENT DATA                    17:09 TUESDAY, MAY 7, 1985    4
              PLOT OF PERSAND*PERGRAV    SYMBOL IS VALUE OF LOCATION
```



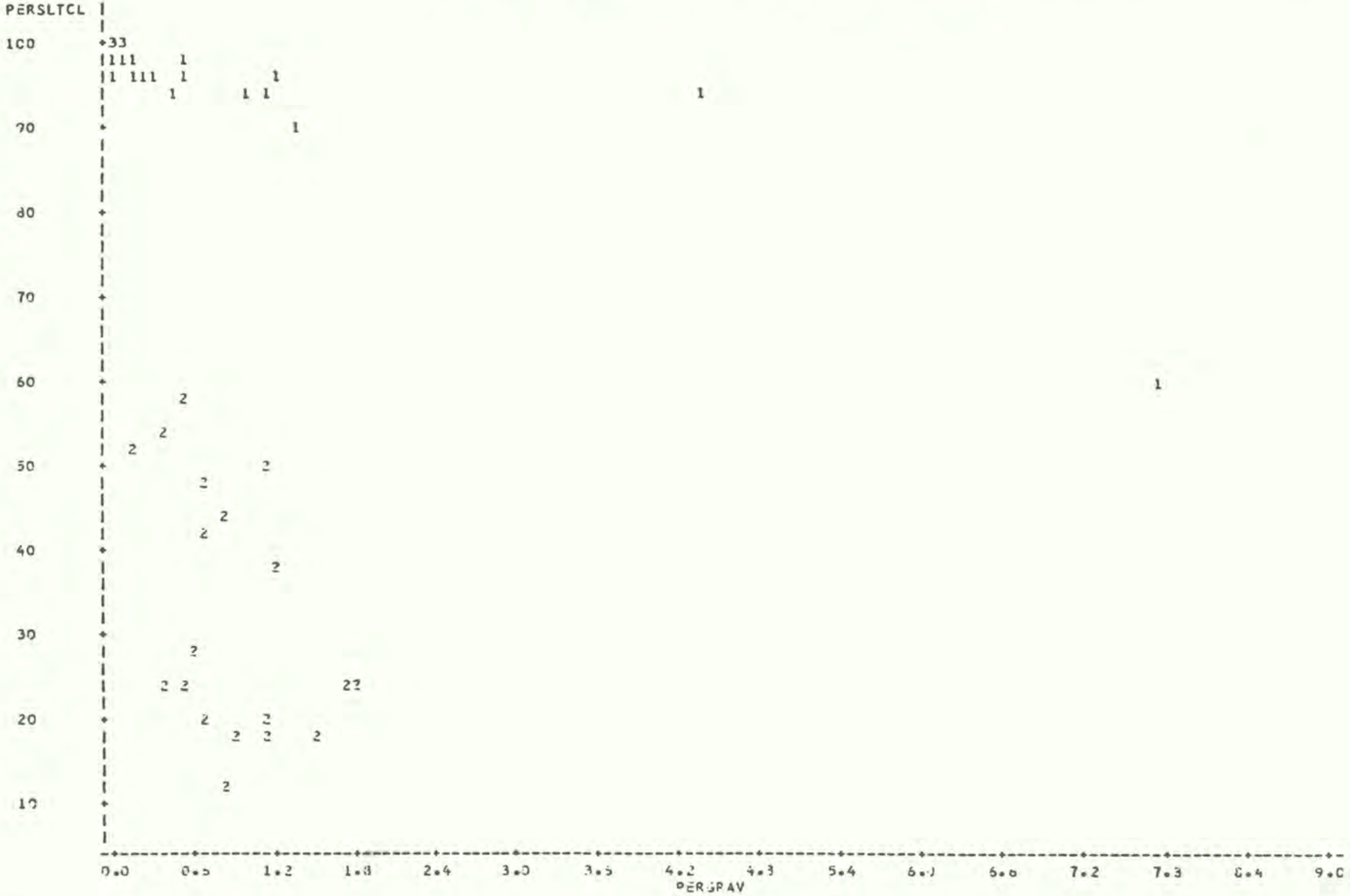
```
NOTE:     23 OBS HIDDEN
```

```
PLOT OF PERSAND*PERORG    SYMBOL IS VALUE OF LOCATION
```



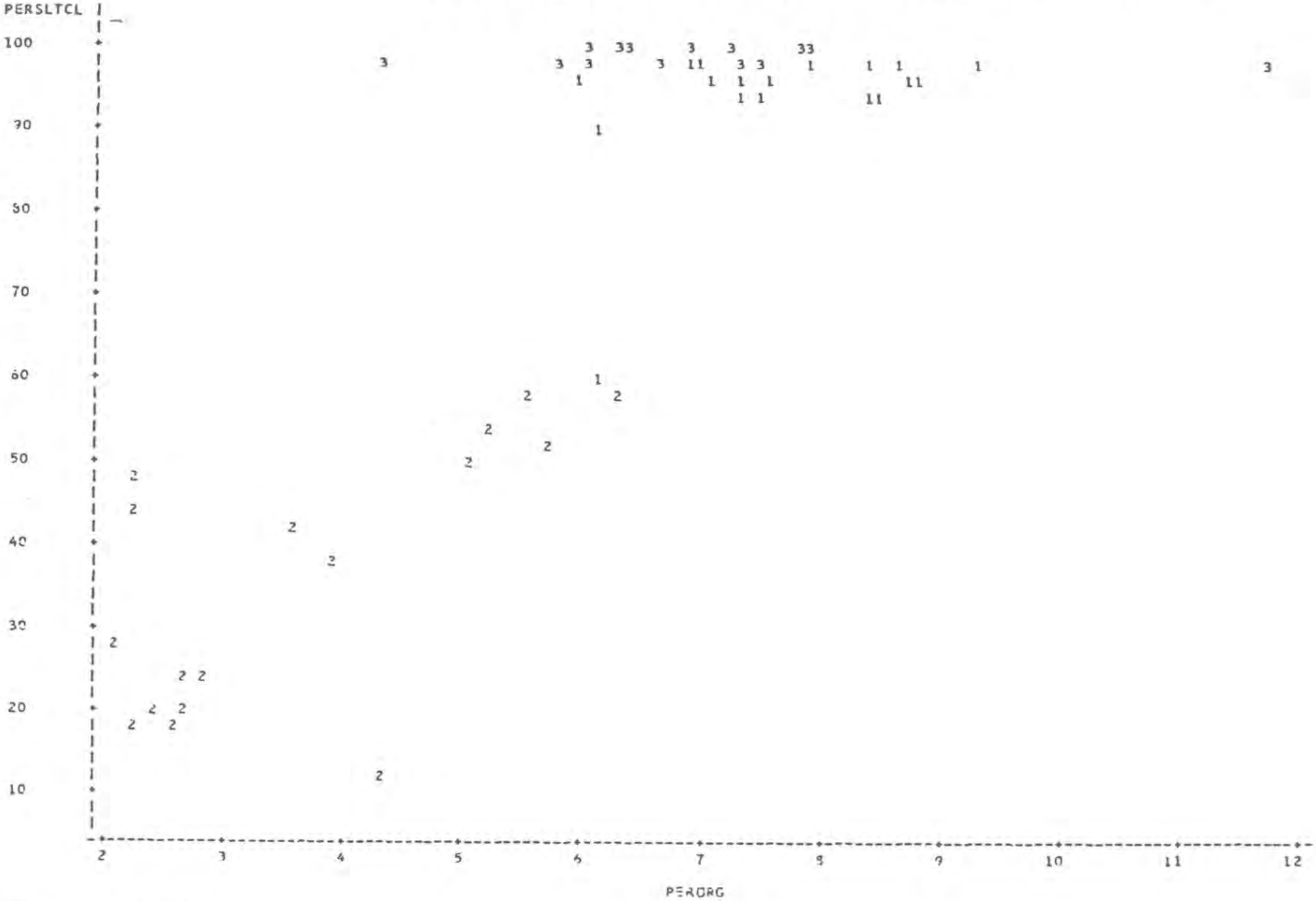
```
NOTE:      7 OBS HIDDEN
```

```
                    SAS ANALYSIS ON SEDIMENT DATA                     17:09 TUESDAY, MAY 7, 1985    6
               PLOT OF PERSLTCL*PERGRAV    SYMBOL IS VALUE OF LOCATION
```



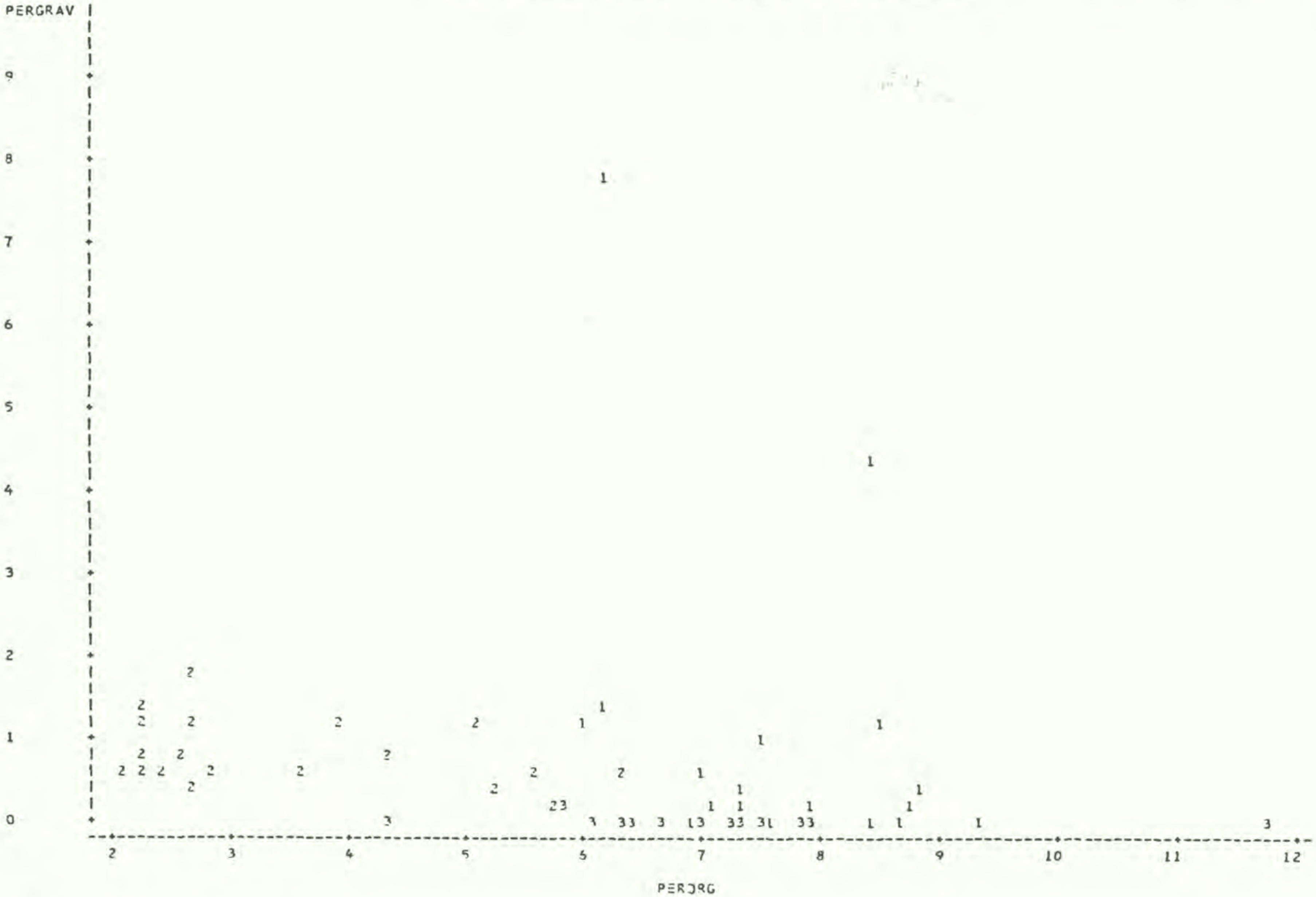
```
NOTE:     23 OBS HIDDEN
```

PLOT OF PERSLTCL*PERORG   SYMBOL IS VALUE OF LOCATION



NOTE: 11 OBS HIDDEN

PLOT OF PERGRAV*PERORG SYMBOL IS VALUE OF LOCATION



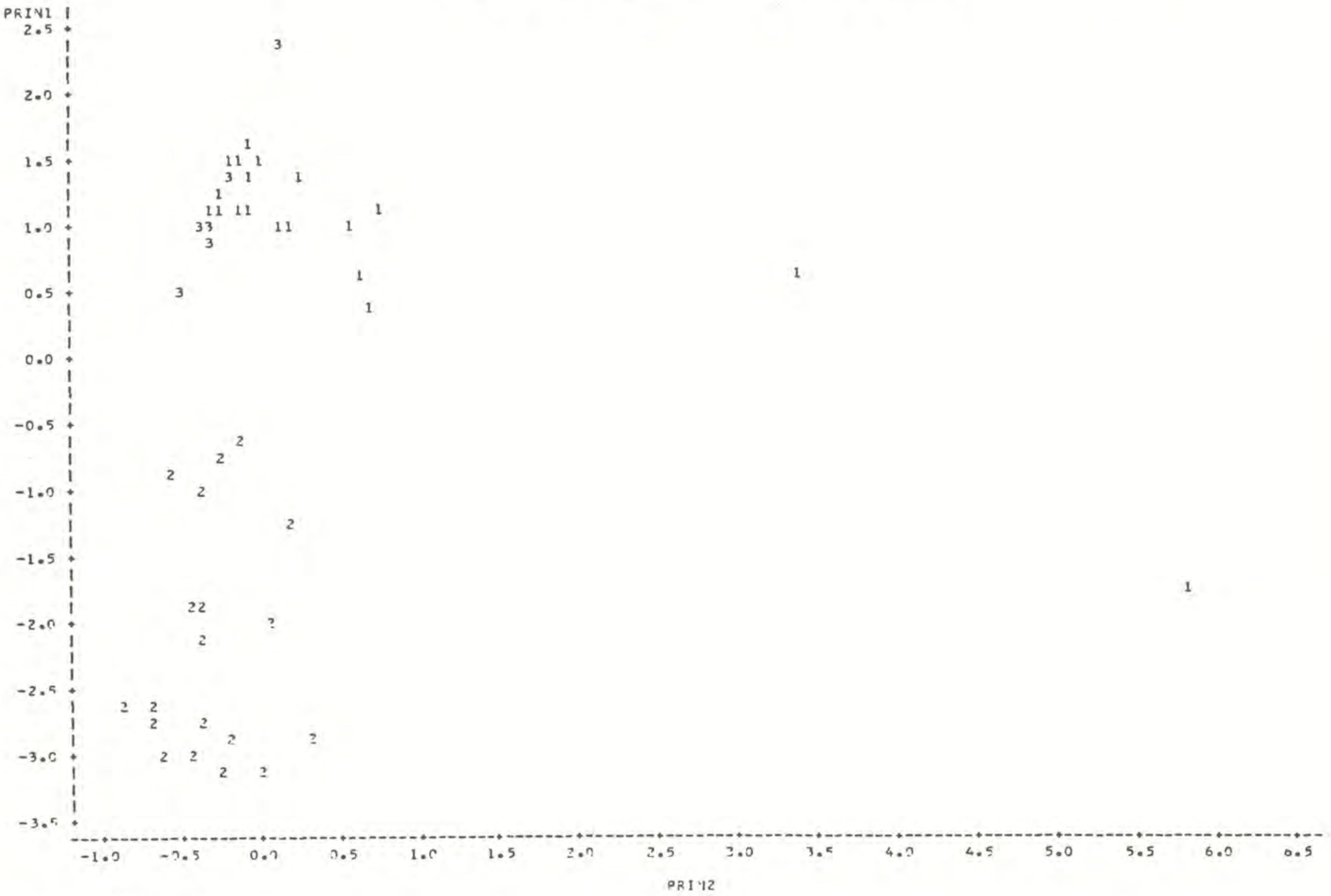
NOTE: 10 OBS HIDDEN

```
                              SAS ANALYSIS ON SEDIMENT DATA                    17:09 TUESDAY, MAY 7, 1985    5
                              PRINCIPAL COMPONENT ANALYSIS

60 OBSERVATIONS
 4 VARIABLES

                                   SIMPLE STATISTICS

                    PERSAND        PERSLTCL         PERGRAV          PERORG

      MEAN         23.93125        75.45463        0.614950        6.079983
      ST DEV       31.21103        31.55470        1.174217        2.260263

                                     COVARIANCES

                    PERSAND        PERSLTCL         PERGRAV          PERORG

      PERSAND        974.13          -984.2          10.077          -60.75
      PERSLTCL       -984.2           995.7          -11.46          61.266
      PERGRAV        10.077          -11.46          1.3788          -.5152
      PERORG         -60.75          61.266          -.5152          5.1088

                          TOTAL VARIANCE =     1976.315

                EIGENVALUE       DIFFERENCE        PROPORTION        CUMULATIVE

 PRIN1            1973.098         1971.150             0.998             0.998
 PRIN2               1.948            0.677             0.001             0.999
 PRIN3               1.270            1.270             0.001             1.000
 PRIN4               0.000              .               0.000             1.000

                                    EIGENVECTORS

                      PRIN1           PRIN2           PRIN3           PRIN4

      PERSAND      -.702525        -.368586        0.148886        0.577315
      PERSLTCL     0.710263        -.393550        0.085707        0.577315
      PERGRAV      -.007729        0.752076        -.234269        0.577421
      PERORG       0.043800        0.237187        0.956373        -.000170
```

```
                               SAS ANALYSIS ON SEDIMENT DATA                    17:09 TUESDAY, MAY 7, 1985   10
                               PRINCIPAL COMPONENT ANALYSIS

60 OBSERVATIONS
 4 VARIABLES

                                        SIMPLE STATISTICS

                      PERSAND        PERSLTCL         PERGRAV         PERORG

         MEAN        23.73125        75.45463        0.614950       6.079983
         ST DEV      31.21103        31.55470        1.174217       2.260263

                                         CORRELATIONS

                      PERSAND        PERSLTCL         PERGRAV         PERORG

         PERSAND       1.0000          -.9994          0.2750          -.8611
         PERSLTCL      -.9994          1.0000          -.3092          0.8590
         PERGRAV       0.2750          -.3092          1.0000          -.1941
         PERORG        -.8611          0.8590          -.1941          1.0000


                 EIGENVALUE      DIFFERENCE      PROPORTION      CUMULATIVE

       PRIN1      2.920733        2.019341        0.730183        0.730183
       PRIN2      0.900892        0.722518        0.225223        0.955406
       PRIN3      0.178375        0.173375        0.044594        1.000000
       PRIN4      0.000000                        0.000000        1.000000

                                         EIGENVECTORS

                        PRIN1           PRIN2           PRIN3           PRIN4

         PERSAND      -.572831        -.116183        0.405131        0.702978
         PERSLTCL     0.575171        0.079856        -.397286        0.710719
         PERGRAV      -.228839        0.968759        -.091744        0.026452
         PERORG       0.537211        0.204422        0.818301        -.000015
```

```
                                                   SAS ANALYSIS ON SEDIMENT DATA                      17:09 TUESDAY, MAY 7, 1985   11
                                              PLOT OF PRIN1*PRIN2    SYMBOL IS VALUE OF LOCATION
PRIN1 |        3     3333333 1           3
 30   +               3 113 111 11    1                                                   1
      |                   1 1         1 1   1
      |                         1
      |
 20   +                                  1
      |
      |
      |
 10   +
      |
      |
      |
  0   +
      |
      |
      |
-10   +
      |
      |                                                                                                                             1
      |
-20   +
      |
      |                      2  2
      |
-30   +
      |                  22
      |
      |             2               2
-40   +
      |
      |                2
      |                  2
-50   +
      |                           2
      |
      |
-60   +
      |
      |
      |             2
-70   +
      |           2   2               22
      |
      |                       2
-80   +                2  2      2
      |                      2
      |
      |                         2
-90   +
      -------+-----------------+-----------------+-----------------+-----------------+-----------------+-----------------+------------
           -1.5               0.0               1.5               3.0               4.5               6.0               7.5

                                                                 PRIN2

NOTE:      11 OBS HIDDEN
```

```
                    SAS ANALYSIS ON SEDIMENT DATA                17:09 TUESDAY, MAY 7, 1985   12
          PLOT OF PRIN1*PRIN2     SYMBOL IS VALUE OF LOCATION
```



NOTE: 16 OBS HIDDEN

```
                          SAS ANALYSIS ON SEDIMENT DATA                         17:09 TUESDAY, MAY 7, 1985   13
                  WARD'S MINIMUM VARIANCE HIERARCHICAL CLUSTER ANALYSIS

                           EIGENVALUES OF THE COVARIANCE MATRIX

                      EIGENVALUE       DIFFERENCE      PROPORTION     CUMULATIVE

                        1973.098         1971.150           0.998          0.998
                           1.948            0.677           0.001          0.999
                           1.270            1.270           0.001          1.000
                           0.000                .           0.000          1.000

                 ROOT-MEAN-SQUARE TOTAL-SAMPLE STANDARD DEVIATION =  22.2279
                 ROOT-MEAN-SQUARE DISTANCE BETWEEN OBSERVATIONS   =  44.4558

  NUMBER      FREQUENCY      RMS STD                                      APPROXIMATE       CUBIC
    OF         OF NEW         OF NEW      SEMIPARTIAL                       EXPECTED      CLUSTERING
 CLUSTERS      CLUSTER        CLUSTER      R-SQUARED        R-SQUARED       R-SQUARED      CRITERION

    10              8        1.57967       0.000335         0.998145        0.992461         7.7405
     9             31       0.927458       0.000383         0.997762        0.990491         8.0046
     8              6        2.02742       0.000597         0.997164        0.987673         8.1599
     7              6        2.70773       0.000976         0.996139        0.983463         8.1911
     6             39        1.47425       0.001349         0.994840        0.976815         8.4546
     5             11        2.93514       0.002092         0.992748        0.965504         8.8987
     4              7        4.24087       0.002444         0.990304        0.944134        10.2639
     3             10        5.85681       0.006683         0.983621        0.896731        11.4793
     2             21        11.5844       0.078526         0.905095        0.757624         7.1540
     1             60        22.2279       0.905095         0.000000        0.000000         0.0000
```

```
                                                    SAS ANALYSIS ON SEDIMENT DATA                        17:09 TUESDAY, MAY 7, 1985   14
OBS      _NAME_         _PARENT_    _NCL_    _FREQ_     _RMSSTD_      _DIST_       _AVLINK_            _SPRSQ_

  1                3     CL59        60        1       0.000000     0.0000000     0.0000000                  0
  2                3     CL59        60        1       0.000000     0.0000000     0.0000000                  0
  3                1     CL58        60        1       0.000000     0.0000000     0.0000000                  0
  4                1     CL58        60        1       0.000000     0.0000000     0.0000000                  0
  5                2     CL57        60        1       0.000000     0.0000000     0.0000000                  0
  6                2     CL57        60        1       0.000000     0.0000000     0.0000000                  0
  7                3     CL56        60        1       0.000000     0.0000000     0.0000000                  0
  8                3     CL56        60        1       0.000000     0.0000000     0.0000000                  0
  9                1     CL55        60        1       0.000000     0.0000000     0.0000000                  0
 10                1     CL55        60        1       0.000000     0.0000000     0.0000000                  0
 11                3     CL54        60        1       0.000000     0.0000000     0.0000000                  0
 12                3     CL54        60        1       0.000000     0.0000000     0.0000000                  0
 13                3     CL53        60        1       0.000000     0.0000000     0.0000000                  0
 14                3     CL53        60        1       0.000000     0.0000000     0.0000000                  0
 15                3     CL52        60        1       0.000000     0.0000000     0.0000000                  0
 16                3     CL52        60        1       0.000000     0.0000000     0.0000000                  0
 17                3     CL51        60        1       0.000000     0.0000000     0.0000000                  0
 18                3     CL51        60        1       0.000000     0.0000000     0.0000000                  0
 19                1     CL50        60        1       0.000000     0.0000000     0.0000000                  0
 20                1     CL50        60        1       0.000000     0.0000000     0.0000000                  0
 21     CL59             CL49        59        2       0.029445     0.0018734     0.0018734        2.97420E-08
 22                3     CL49        60        1       0.000000     0.0000000     0.0000000                  0
 23                3     CL48        60        1       0.000000     0.0000000     0.0000000                  0
 24     CL53             CL48        53        2       0.100067     0.0063666     0.0063666        3.43594E-07
 25                1     CL47        60        1       0.000000     0.0000000     0.0000000                  0
 26     CL55             CL47        55        2       0.076021     0.0048367     0.0048367        1.98250E-07

OBS      _RSQ_       _ERSQ_      _RATIO_     _LOGR_        _CCC_      PERSAND     PERSLTCL     PERGRAV      PERORG      LOCATION

  1     1.00000        .            .           .             .         0.6500      99.3500      0.00000     7.90800         3
  2     1.00000        .            .           .             .         0.6000      99.4000      0.00000     7.86400         3
  3     1.00000        .            .           .             .         2.5200      97.4800      0.00000     6.87900         1
  4     1.00000        .            .           .             .         2.4600      97.4670      0.07300     6.89400         1
  5     1.00000        .            .           .             .        74.6450      23.5670      1.78800     2.63800         2
  6     1.00000        .            .           .             .        74.6990      23.5810      1.72000     2.70500         2
  7     1.00000        .            .           .             .         2.1930      97.8070      0.00000     7.53500         3
  8     1.00000        .            .           .             .         2.0780      97.9220      0.00000     7.52600         3
  9     1.00000        .            .           .             .         2.8650      96.9630      0.17200     8.73400         1
 10     1.00000        .            .           .             .         2.9530      97.0470      0.00000     8.69100         1
 11     1.00000        .            .           .             .         1.4690      98.5310      0.00000     7.52400         3
 12     1.00000        .            .           .             .         1.6530      98.3470      0.00000     7.49000         3
 13     1.00000        .            .           .             .         0.9600      99.0050      0.03500     6.39600         3
 14     1.00000        .            .           .             .         1.0870      98.9130      0.00000     6.62900         3
 15     1.00000        .            .           .             .         0.7540      99.1890      0.05800     5.33300         3
 16     1.00000        .            .           .             .         0.8440      99.1260      0.03000     6.06800         3
 17     1.00000        .            .           .             .         1.8320      98.1680      0.00000     7.29800         3
 18     1.00000        .            .           .             .         1.8010      98.1780      0.02100     6.79700         3
 19     1.00000        .            .           .             .         4.2370      95.5930      0.17000     7.30300         1
 20     1.00000        .            .           .             .         4.3290      95.6710      0.00000     7.56000         1
 21     1.00000     1.00000     14.3934     2.66609       25.2623       0.6250      99.3750      0.00000     7.88600         .
 22     1.00000        .            .           .             .         0.8550      99.1450      0.00000     7.91400         3
 23     1.00000        .            .           .             .         0.9850      99.0150      0.00000     6.89500         3
 24     1.00000     0.99999     21.0662     3.04757       28.9799       1.0235      98.9590      0.01750     6.51250         .
 25     1.00000        .            .           .             .         3.1170      96.6920      0.19100     8.75900         1
 26     1.00000     0.99999     25.9800     3.25347       30.9285       2.9090      97.0050      0.08600     8.71250         .
```

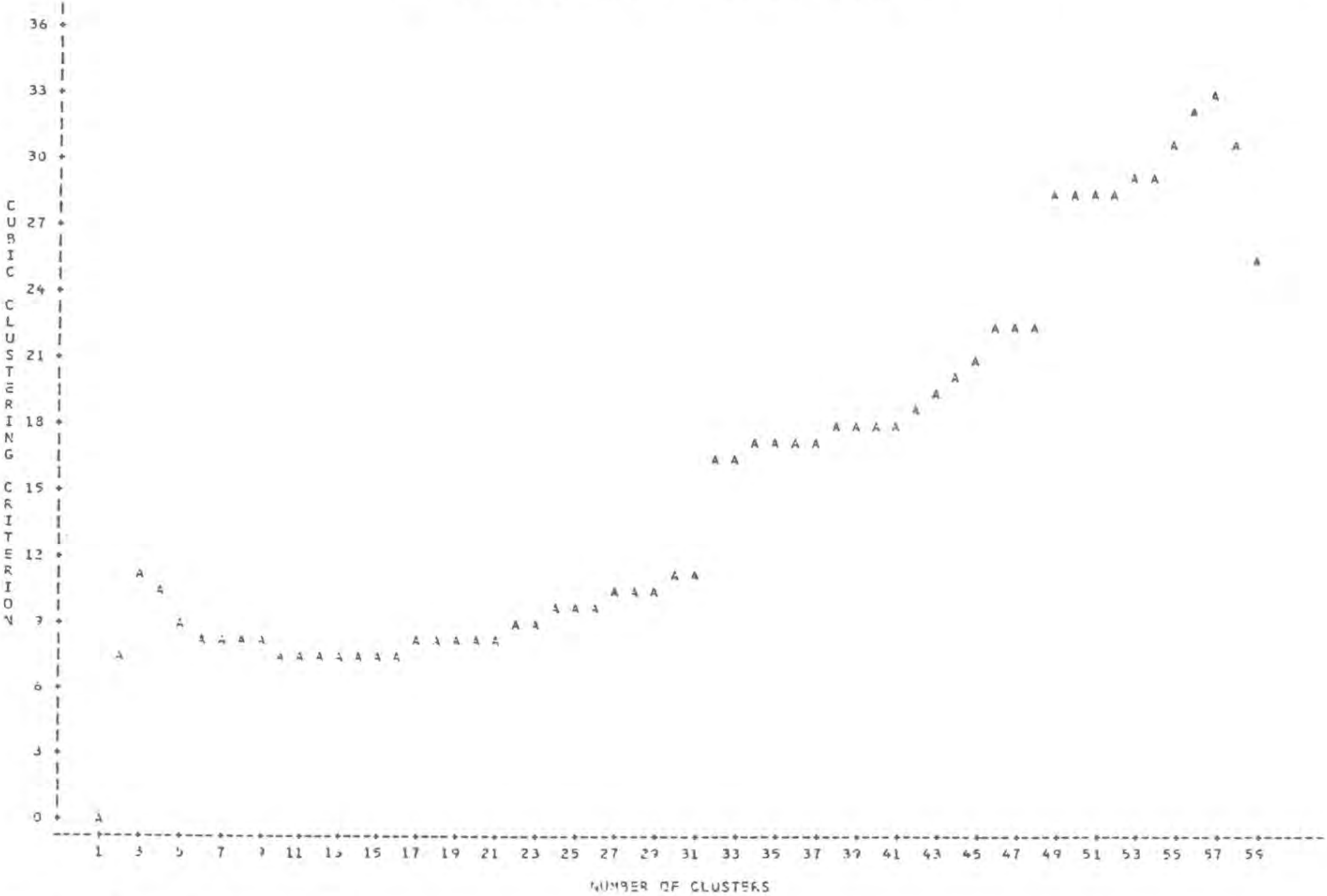
```
                                              SAS ANALYSIS ON SEDIMENT DATA                             17:09 TUESDAY, MAY 7, 1985   15
OBS       _NAME_        _PARENT_     _NCL_     _FREQ_     _RMSSTO_      _DIST_         _AVLINK_        _SPRSQ_

 27                 3     CL46         60         1       0.000000     0.0000000      0.0000000      0.0000000000
 28      CL52             CL46         52         2       0.101827     0.0064787      0.0064787      0.0000003557
 29                 1     CL45         60         1       0.000000     0.0000000      0.0000000      0.0000000000
 30      CL54             CL45         54         2       0.092782     0.0059031      0.0059031      0.0000002953
 31      CL58             CL44         58         2       0.034138     0.0021719      0.0021719      0.0000000400
 32                 1     CL44         60         1       0.000000     0.0000000      0.0000000      0.0000000000
 33                 2     CL43         60         1       0.000000     0.0000000      0.0000000      0.0000000000
 34                 2     CL43         60         1       0.000000     0.0000000      0.0000000      0.0000000000
 35      CL49             CL42         49         3       0.096517     0.0073438      0.0074032      0.0000006094
 36                 3     CL42         60         1       0.000000     0.0000000      0.0000000      0.0000000000
 37                 1     CL41         60         1       0.000000     0.0000000      0.0000000      0.0000000000
 38      CL47             CL41         47         3       0.125500     0.0086368      0.0091617      0.0000009824
 39      CL56             CL40         56         2       0.057598     0.0036639      0.0036639      0.0000001138
 40      CL51             CL40         51         2       0.107293     0.0068267      0.0068267      0.0000003949
 41                 2     CL39         60         1       0.000000     0.0000000      0.0000000      0.0000000000
 42                 2     CL39         60         1       0.000000     0.0000000      0.0000000      0.0000000000
 43      CL46             CL38         46         3       0.138634     0.0092314      0.0097832      0.0000009629
 44      CL48             CL38         48         3       0.132700     0.0087477      0.0093089      0.0000008647
 45      CL44             CL37         44         3       0.198147     0.0153251      0.0153635      0.0000026538
 46                 1     CL37         60         1       0.000000     0.0000000      0.0000000      0.0000000000
 47      CL45             CL36         45         3       0.175870     0.0127150      0.0130531      0.0000018268
 48      CL40             CL36         40         4       0.183269     0.0131881      0.0137453      0.0000029479
 49                 1     CL35         60         1       0.000000     0.0000000      0.0000000      0.0000000000
 50                 1     CL35         60         1       0.000000     0.0000000      0.0000000      0.0000000000
 51                 2     CL34         60         1       0.000000     0.0000000      0.0000000      0.0000000000
 52                 2     CL34         60         1       0.000000     0.0000000      0.0000000      0.0000000000

OBS      _RSQ_      _ERSQ_      _RATIO_      _LOGR_       _CCC_       PERSAND     PERSLTCL      PERGRAV      PERORG       LOCATION

 27     1.00000       .            .           .            .          0.5500      99.4500      0.00000      6.33700          3
 28     1.00000    0.99997      20.9430     3.04130      29.8235       0.7990      99.1570      0.04400      6.20050          .
 29     1.00000       .            .           .            .          1.7610      98.1230      0.16600      7.89700          1
 30     1.00000    0.99998      22.4583     3.11217      29.4892       1.5610      98.4390      0.00000      7.50700          .
 31     1.00000    1.00000      24.8404     3.21247      30.4397       2.4900      97.4735      0.03650      6.98650          .
 32     1.00000       .            .           .            .          2.0960      97.3440      0.55900      7.03400          1
 33     1.00000       .            .           .            .         80.3880      18.1460      1.46600      2.25900          2
 34     1.00000       .            .           .            .         80.5340      18.5720      0.89400      2.58400          2
 35     1.00000    0.99994      20.3204     3.01162      29.5185       0.7017      99.2983      0.00000      7.89533          .
 36     1.00000       .            .           .            .          0.7040      99.2960      0.00000      7.22800          3
 37     1.00000       .            .           .            .          2.8500      97.1500      0.00000      9.32500          1
 38     1.00000    0.99992      18.2251     2.90230      22.4603       2.1763      97.7027      0.12100      8.72767          .
 39     1.00000    0.99999      30.0937     3.40449      37.2593       2.1355      97.8645      0.00000      7.53050          .
 40     1.00000    0.99996      21.0503     3.04694      29.8725       1.8165      98.1730      0.01050      7.14750          .
 41     1.00000       .            .           .            .         42.2570      57.2070      0.53600      5.54200          2
 42     1.00000       .            .           .            .         41.9570      57.4940      0.54900      6.29900          2
 43     0.99999    0.99990      17.7694     2.87748      22.2648       0.7160      99.2547      0.02933      6.24600          .
 44     1.00000    0.99993      18.8949     2.93889      22.7392       1.0107      98.9777      0.01167      6.64000          .
 45     0.99999    0.99987      13.3097     2.58847      20.0290       2.3593      97.4303      0.21033      6.93567          .
 46     1.00000       .            .           .            .          2.8940      96.8200      0.28500      7.05800          1
 47     0.99999    0.99989      15.6242     2.74882      21.2677       1.6277      98.3177      0.05533      7.63700          .
 48     0.99999    0.99978      10.4317     2.34485      18.1462       1.9750      98.0137      0.00525      7.33900          .
 49     1.00000       .            .           .            .          4.6220      94.2500      1.12800      8.47900          1
 50     1.00000       .            .           .            .          4.3860      94.6200      0.99400      7.53000          1
 51     1.00000       .            .           .            .         46.6070      53.0030      0.39000      5.23800          2
 52     1.00000       .            .           .            .         47.4140      52.4180      0.16800      5.70900          2
```

```
                                                 SAS ANALYSIS ON SEDIMENT DATA                         17:09 TUESDAY, MAY 7, 1985   16
OBS       _NAME_           _PARENT_      _NCL_     _FREQ_      _RMSSTD_       _DIST_         _AVLINK_        _SPRSQ_          _RSQ_

 53                   2      CL33         60          1        0.00000       0.000000       0.000000       0.000000000      1.00000
 54      CL43                CL33         43          2        0.28187       0.017933       0.017933       0.000002726      0.99999
 55                   3      CL32         60          1        0.00000       0.000000       0.000000       0.000000000      1.00000
 56                   3      CL32         60          1        0.00000       0.000000       0.000000       0.000000000      1.00000
 57      CL41                CL31         41          4        0.19655       0.015091       0.015779       0.000002895      0.99998
 58                   1      CL31         60          1        0.00000       0.000000       0.000000       0.000000000      1.00000
 59                   2      CL30         60          1        0.00000       0.000000       0.000000       0.000000000      1.00000
 60                   2      CL30         60          1        0.00000       0.000000       0.000000       0.000000000      1.00000
 61      CL35                CL29         35          2        0.37269       0.023711       0.023711       0.000004765      0.99996
 62                   1      CL29         60          1        0.00000       0.000000       0.000000       0.000000000      1.00000
 63      CL31                CL28         31          5        0.29781       0.024583       0.025748       0.000008194      0.99993
 64                   1      CL28         60          1        0.00000       0.000000       0.000000       0.000000000      1.00000
 65      CL37                CL27         37          4        0.26195       0.018537       0.019914       0.000004368      0.99997
 66      CL32                CL27         32          2        0.45404       0.028888       0.028888       0.000007072      0.99994
 67      CL33                CL26         33          3        0.36206       0.023553       0.025203       0.000006269      0.99995
 68                   2      CL26         60          1        0.00000       0.000000       0.000000       0.000000000      1.00000
 69      CL30                CL25         30          2        0.55233       0.035173       0.035173       0.000010484      0.99992
 70      CL57                CL25         57          2        0.03909       0.002487       0.002487       0.000000052      1.00000
 71      CL50                CL24         50          2        0.11699       0.007443       0.007443       0.000000470      1.00000
 72                   1      CL24         60          1        0.00000       0.000000       0.000000       0.000000000      1.00000
 73      CL42                CL23         42          4        0.18451       0.015011       0.015424       0.000002865      0.99998
 74      CL38                CL23         38          6        0.19663       0.012707       0.014531       0.000004105      0.99997
 75      CL27                CL22         27          6        0.42943       0.027825       0.032970       0.000017497      0.99987
 76      CL36                CL22         36          7        0.22221       0.012561       0.015827       0.000004584      0.99996
 77                   2      CL21         60          1        0.00000       0.000000       0.000000       0.000000000      1.00000
 78                   2      CL21         60          1        0.00000       0.000000       0.000000       0.000000000      1.00000

OBS      _ERSQ_      _RATIO_     _LOGR_        _CCC_        PERSAND      PERSLTCL      PERGRAV       PERORG     LOCATION

 53        .            .           .             .         80.0960      19.2110      0.69300       2.4120         2
 54     0.99985      12.0217     2.48571       19.2424      80.4610      18.3590      1.13000       2.4215         .
 55        .            .           .             .          1.6680      98.3320      0.00000       6.0860         3
 56        .            .           .             .          2.5040      97.3910      0.10500       5.8540         3
 57     0.99930      10.7159     2.37182       18.3544       2.9463      96.9630      0.09075       8.8770         .
 58        .            .           .             .          2.3090      97.6910      0.00000       8.3770         1
 59        .            .           .             .         74.8740      24.5880      0.53800       2.8560         2
 60        .            .           .             .         76.0560      23.6020      0.34200       2.6630         2
 61     0.99961       9.1094     2.20931       17.1005       4.5040      94.4350      1.06100       8.0045         .
 62        .            .           .             .          5.6950      93.8410      0.46400       7.3300         1
 63     0.99943       8.2129     2.10570       11.5274       2.8188      97.1086      0.07260       8.7770         .
 64        .            .           .             .          3.5640      95.9440      0.49200       8.8400         1
 65     0.99969       9.4838     2.24954       17.4109       2.4930      97.2777      0.22925       6.9662         .
 66     0.99943       8.4859     2.13353       16.5554       2.0860      97.8615      0.05250       5.9700         .
 67     0.99953       8.7130     2.16492       16.7530      80.3393      18.6430      1.01767       2.4183         .
 68        .            .           .             .         81.4490      17.4010      1.15000       2.2360         2
 69     0.99937       7.8209     2.05679       11.2604      75.4650      24.0950      0.44000       2.7595         .
 70     1.00000      32.2959     3.47494       32.7268      74.6720      23.5740      1.75400       2.6715         .
 71     0.99995      20.9569     3.04247       29.5305       4.2830      95.6320      0.08500       7.4315         .
 72        .            .           .             .          3.2950      95.4770      1.22800       5.9720         1
 73     0.99992      11.2003     2.41594       18.6953       0.7023      99.2977      0.00000       7.7295         .
 74     0.99972       9.7998     2.28235       17.6539       0.8633      99.1162      0.02050       6.4430         .
 75     0.99918       6.4754     1.86801       10.2293       2.3573      97.4723      0.17033       6.5342         .
 76     0.99965       9.2531     2.22550       17.2252       1.3257      98.1537      0.02671       7.4667         .
 77        .            .           .             .         49.4750      49.3950      1.12900       5.0730         2
 78        .            .           .             .         50.6990      48.6190      0.68200       2.2750         2
```

```
OBS      _NAME_         _PARENT_     _NCL_    _FREQ_     _RMSSTD_     _DIST_      _AVLINK_      _SPRSQ_      _RSQ_

 79     CL29              CL20          29        3        0.5335    0.03615     0.03804      0.000015    0.99991
 80     CL24              CL20          24        3        0.6137    0.04738     0.04753      0.000025    0.99981
 81                2      CL19          60        1        0.0000    0.00000     0.00000      0.000000    1.00000
 82                2      CL19          60        1        0.0000    0.00000     0.00000      0.000000    1.00000
 83     CL23              CL18          23       10        0.3833    0.02943     0.03135      0.000035    0.99977
 84                3      CL18          60        1        0.0000    0.00000     0.00000      0.000000    1.00000
 85     CL26              CL17          26        4        0.5137    0.03731     0.04008      0.000018    0.99985
 86                2      CL17          60        1        0.0000    0.00000     0.00000      0.000000    1.00000
 87     CL28              CL16          28        6        0.3976    0.03253     0.03467      0.000015    0.99989
 88                3      CL16          60        1        0.0000    0.00000     0.00000      0.000000    1.00000
 89     CL25              CL15          25        4        0.5674    0.03651     0.04055      0.000023    0.99983
 90                2      CL15          60        1        0.0000    0.00000     0.00000      0.000000    1.00000
 91     CL19              CL14          19        2        1.2223    0.07777     0.07777      0.000061    0.99959
 92                2      CL14          60        1        0.0000    0.00000     0.00000      0.000000    1.00000
 93     CL20              CL13          20        6        0.7288    0.04242     0.05188      0.000046    0.99964
 94                1      CL13          60        1        0.0000    0.00000     0.00000      0.000000    1.00000
 95     CL22              CL12          22       13        0.4470    0.02718     0.03369      0.000040    0.99973
 96     CL18              CL12          18       11        0.5473    0.06113     0.06328      0.000058    0.99953
 97     CL34              CL11          34        2        0.3976    0.02530     0.02530      0.000005    0.99993
 98     CL21              CL11          21        2        1.1253    0.07160     0.07160      0.000043    0.99969
 99     CL13              CL10          13        7        1.1333    0.10920     0.11323      0.000173    0.99888
100                1      CL10          60        1        0.0000    0.00000     0.00000      0.000000    1.00000
101     CL16              CL9           16        7        0.8349    0.08949     0.09097      0.000116    0.99934
102     CL12              CL9           12       24        0.6741    0.04143     0.05140      0.000173    0.99871
103     CL17              CL8           17        5        0.8864    0.07712     0.07968      0.000081    0.99945
104                2      CL8           60        1        0.0000    0.00000     0.00000      0.000000    1.00000

OBS     _ERSQ_       _RATIO_     _LOGR_       _CCC_        PERSAND     PERSLTCL      PERGRAV       PERORG     LOCATION

 79     0.999315      7.2338     1.97877     10.8341       4.9010     94.2370      0.86200       7.7797         .
 80     0.998907      5.6574     1.73296      9.4929       3.9537     95.5903      0.46600       6.9450         .
 81        .            .           .            .         58.2620     41.0690      0.67000       3.5920         2
 82        .            .           .            .         60.3820     38.4120      1.20600       3.9330         2
 83     0.998793      5.2800     1.66393      9.1160       0.7989     99.1888      0.01230       6.9572         .
 84        .            .           .            .          1.2030     98.7920      0.00300       4.3000         3
 85     0.999036      6.2186     1.82755     10.0087      30.6157     18.3325      1.05075       2.3728         .
 86        .            .           .            .         78.1500     20.6900      1.16000       2.6899         2
 87     0.999249      6.8474     1.92397     10.5343       2.9430     96.9145      0.14250       8.7375         .
 88        .            .           .            .          1.1500     98.8500      0.00000      11.7620         3
 89     0.999007      5.9143     1.77746      9.7355      75.0635     23.8345      1.09700       2.7155         .
 90        .            .           .            .         72.3710     27.0430      0.58600       2.0550         2
 91     0.998150      4.5183     1.50813      9.2684      57.3220     39.7400      0.93300       3.7625         .
 92        .            .           .            .         55.8700     43.3140      0.81600       2.2430         2
 93     0.998343      4.6125     1.52878      8.3300       4.4273     94.9087      0.66400       7.3623         .
 94        .            .           .            .          1.5130     94.1700      4.31700       8.4510         1
 95     0.998665      4.9639     1.60217      9.7790       2.0716     97.8392      0.09300       7.0825         .
 96     0.997917      4.4596     1.49507      8.1995       0.8361     99.1527      0.01118       6.7156         .
 97     0.999572      8.9311     2.18954     16.9435      47.0105     52.7105      0.27900       5.4735         .
 98     0.998518      4.7444     1.55696      8.5327      50.0870     49.0375      0.90550       3.6740         .
 99     0.995733      3.7932     1.33320      7.3307       4.0110     94.8031      1.18596       7.5179         .
100        .            .           .            .          8.5610     90.0980      1.34100       5.1720         1
101     0.997307      4.0557     1.40013      7.5845       2.6869     97.1910      0.12214       9.2124         .
102     0.994951      3.9130     1.36429      7.5090       1.5053     98.4412      0.05550       6.9143         .
103     0.997641      4.3096     1.46061      9.0132      80.1234     18.8040      1.07260       2.4360         .
104        .            .           .            .         86.5580     12.6070      0.83500       4.3500         2
```

```
                                           SAS ANALYSIS ON SEDIMENT DATA                      17:09 TUESDAY, MAY 7, 1985   18
OBS      _NAME_        _PARENT_    _NCL_    _FREQ_    _RMSSTD_     _DIST_     _AVLINK_     _SPRSQ_       _RSQ_

105    CL11              CL7         11         4       1.6453    0.11647    0.12250     0.000230     0.99848
106    CL39              CL7         39         2       0.3029    0.01927    0.01927     0.000003     0.99998
107    CL9               CL6          9        31       0.9275    0.06459    0.07914     0.000383     0.99776
108    CL10              CL6         10         8       1.5797    0.15026    0.15750     0.000335     0.99814
109    CL15              CL5         15         5       1.0746    0.09614    0.09865     0.000125     0.99921
110    CL8               CL5          8         6       2.0274    0.20563    0.20870     0.000597     0.99716
111              I       CL4         60         1       0.0000    0.00000    0.00000     0.000000     1.00000
112    CL7               CL4          7         6       2.7077    0.20778    0.21766     0.000976     0.99619
113    CL4               CL3          4         7       4.2409    0.41018    0.42498     0.002444     0.99030
114    CL14              CL3         14         3       1.7315    0.11691    0.12321     0.000154     0.99906
115    CL3               CL2          3        10       5.8568    0.43331    0.47224     0.006683     0.98362
116    CL5               CL2          5        11       2.9351    0.21273    0.23250     0.002092     0.99275
117    CL6               CL1          6        39       1.4742    0.11187    0.13645     0.001349     0.99484
118    CL2               CL1          2        21      11.5844    0.94047    0.98124     0.078526     0.90510
119    CL1                            1        60      22.2279    1.97791    2.04331     0.905095     0.00000

OBS     _ERSQ_       _RATIO_     _LOGR_       _CCC_      PERSAND     PERSLTCL     PERGRAV      PERORG      LOCATION

105    0.993837      4.0206     1.39144      7.6682     48.5487     50.8590     0.59225     4.57375         .
106    0.999748     10.2192     2.32427     17.9378     42.1120     57.3455     0.54250     5.92050         .
107    0.990471      4.2483     1.44652      8.0046      1.7721     98.1589     0.07055     7.43326         .
108    0.992461      4.0539     1.40213      7.7405      4.5798     94.2150     1.20525     7.35212         .
109    0.996900      3.9271     1.36790      7.5112     74.5290     24.4762     0.99480     2.53340         .
110    0.987673      4.3473     1.46956      8.1599     81.1958     17.7712     1.03300     2.75567         .
111        .            .           .            .      31.5430     60.7020     7.75000     6.19000         1
112    0.993463      4.3390     1.46764      8.1911     46.4032     53.0212     0.57567     5.02267         .
113    0.944134      5.7617     1.75123     10.2639     44.2810     54.1184     1.60057     5.18943         .
114    0.996397      3.8180     1.33974      7.3510     58.1713     40.9313     0.89733     3.25500         .
115    0.896731      6.3049     1.84133     11.4793     48.4431     50.1523     1.38960     4.60940         .
116    0.965504      4.7569     1.55959      8.8987     78.1655     20.8189     1.01364     2.67791         .
117    0.976815      4.4933     1.50259      8.4546      2.3481     97.3499     0.30331     7.41562         .
118    0.757624      2.5539     0.93761      7.1540     64.0143     34.7920     1.19371     3.59757         .
119    0.000000      1.0000     0.00000      0.0000     23.9312     75.4546     0.61495     6.07998         .
```

PLOT OF _CCC_*_NCL_   LEGEND: A = 1 OBS, B = 2 OBS, ETC.



NOTE: 60 OBS HAD MISSING VALUES

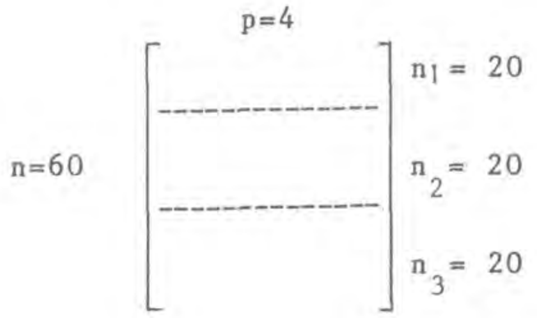
```
EXAMPLE OF SELECTING THE "BEST" VARIABLES FROM A MULTIVARIATE DATA SET

 N =  60        P =    4        CYCLE =  4        BRIEF = 1        CPCT =100.

THE N SAMPLES-BY-P VARIABLES INPUT DATA ARE:

      %SAND  %SILT-CLAY  %GRAVEL     %O.M.
      2.850    97.150     0.000      9.325
      4.622    94.250     1.123      8.479
      3.564    95.944     0.492      8.840
      3.117    96.692     0.191      9.758
      1.513    94.170     4.317      8.451
     46.607    53.003     0.390      5.238
     74.874    24.588     0.538      2.856
     80.096    19.211     0.693      2.412
     81.449    17.401     1.150      2.236
     78.150    20.690     1.160      2.689
      0.650    99.350     0.000      7.908
      0.600    99.400     0.000      7.964
      0.550    99.450     0.000      6.337
      0.985    99.015     0.000      6.895
      0.704    99.296     0.000      7.228
      2.520    97.480     0.000      6.879
      2.894    96.820     0.286      7.058
      2.460    97.467     0.073      6.894
     31.548    60.702     7.750      6.190
      5.895    93.641     0.464      7.330
     49.475    49.396     1.129      5.073
     47.414    52.418     0.168      5.709
     86.558    12.607     0.835      4.360
     50.699    48.619     0.682      2.275
     80.388    18.146     1.466      2.259
      1.208    98.792     0.000      4.300
      1.469    98.531     0.000      7.524
      0.855    99.145     0.000      7.914
      2.193    97.807     0.000      7.535
      1.832    98.166     0.000      7.293
      2.865    96.763     0.172      3.734
      8.561    90.098     1.341      6.192
      4.237    95.593     0.170      7.303
      4.329    95.671     0.000      7.560
      2.098    97.344     0.558      7.034
     58.262    41.068     0.670      3.592
     60.382    38.412     1.206      3.933
     55.870    43.314     0.816      2.243
     74.645    23.567     1.791      2.639
     74.697    23.531     1.720      2.705
      0.960    99.005     0.035      6.396
      0.754    99.193     0.053      6.333
      1.150    98.850     0.000     11.762
      2.078    97.922     0.000      7.526
      1.653    98.347     0.000      7.490
      1.761    98.123     0.165      7.697
      2.309    97.691     0.000      6.377
      %SAND  %SILT-CLAY  %GRAVEL     %O.M.
      4.386    94.620     0.994      7.530
      3.295    95.477     1.228      5.972
      2.953    97.047     0.000      3.691
     42.257    57.207     0.536      5.542
     41.967    57.484     0.549      5.299
     72.371    27.043     0.586      2.055
     76.056    23.602     0.342      2.663
     80.534    18.572     0.894      2.594
      0.844    99.126     0.030      6.063
      1.668    98.332     0.000      5.986
      1.301    98.678     0.021      6.997
      1.087    98.913     0.000      6.629
      2.504    97.391     0.105      3.954

THE P-BY-P CORRELATION MATRIX IS:
  1.0000 -0.9993  0.2750 -0.8611
 -0.9993  1.0000 -0.3092  0.8590
  0.2750 -0.3092  1.0000 -0.1941
 -0.8611  0.8590 -0.1941  1.0000

THE VARIABLES (TOP) ARE ORDERED BY THE SUM-OF-RSQUARED    CRITERION (BOTTOM) :
         2          1          4          3
    2.8322     2.8153     2.5171     1.2089

   1 VARIABLES & THEIR CORRELATIONS HAVE BEEN REMOVED.


THE P-BY-P CORRELATION MATRIX IS:
  0.0013  0.0000 -0.0340 -0.0027
  0.0000  0.0000  0.0000  0.0000
 -0.0340  0.0000  0.9044  0.0715
 -0.0027  0.0000  0.0715  0.2621

THE VARIABLES (TOP) ARE ORDERED BY THE SUM-OF-RSQUARED    CRITERION (BOTTOM) :
         3          4          1          2
    0.9242     0.0733     0.0012     0.0000

   2 VARIABLES & THEIR CORRELATIONS HAVE BEEN REMOVED.


THE P-BY-P CORRELATION MATRIX IS:
  0.0000  0.0000 -0.0000  0.0000
  0.0000  0.0000  0.0000  0.0000
 -0.0000  0.0000  0.0000  0.0000
  0.0000  0.0000  0.0000  0.2564

THE VARIABLES (TOP) ARE ORDERED BY THE SUM-OF-RSQUARED    CRITERION (BOTTOM) :
         4          1          3          2
    0.0658     0.0000     0.0000     0.0000

   3 VARIABLES & THEIR CORRELATIONS HAVE BEEN REMOVED.
```

```
THE P-BY-P CORRELATION MATRIX IS:
  0.0000  0.0000 -0.0000  0.0000
  0.0000  0.0000  0.0000  0.0000
 -0.0000  0.0000  0.0000  0.0000
  0.0000  0.0000  0.0000  0.0000

THE VARIABLES (TOP) ARE ORDERED BY THE SUM-OF-RSQUARED    CRITERION (BOTTOM) :
         1          3          4          2
    0.0000     0.0000     0.0000     0.0000

   4 VARIABLES & THEIR CORRELATIONS HAVE BEEN REMOVED.


THE CORRELATION MATRIX IS NOW EXHAUSTED.

THE BEST ORDER IN WHICH TO CHOOSE VARIABLES, FOR MAXIMUM INFORMATION ABOUT STRUCTURE IN THE DATA SET, IS:
       2        3        4        1

THE TRACES ASSOCIATED WITH THE RESIDUAL CORRELATION MATRICES ARE (BEGINNING WITH 0 VARIABLES REMOVED):
   4.000   1.168   0.255   0.000

THE TRACES, AS A PERCENTAGE OF P, ARE:
   100.0    29.2     6.4     0.0
```

8. MULTIVARIATE ANALYSIS OF VARIANCE (MANOVA) AND DISCRIMINANT ANALYSIS (DA)

8.1 Introduction

We will use the data set 'SEDABC' again but this time our analysis will be based on the *a priori* partitioning of the samples into 3 groups (the 3 locations). Thus the data matrix leads to the MANOVA and DA



We will do the analysis in APL and MINITAB.
In principle MANOVA and DA is simple:

(1) We calculate the deviation squares and cross-products matrix for each of the 3 groups, and then we sum them (matrix addition). That is, we calculate W_1, W_2 and W_3 and then $W=W_1 + W_2 + W_3$. Then we calculate the deviation squares and cross products matrix for all the data regardless of group membership, to get the T matrix. Then the among-group matrix A is obtained by A=T-W.

(2) We can think of a ANOVA table, as in the univariate ANOVA:

Source	df	SS	MS
Among groups	g-1	$A= P\overset{P}{\begin{bmatrix} \quad & \quad \end{bmatrix}}$	Among group pxp covariance matrix
Within groups	n-g	$W= P\overset{P}{\begin{bmatrix} \quad & \quad \end{bmatrix}}$	Within group pxp covariance matrix
Total	n-1	$T= P\overset{P}{\begin{bmatrix} \quad & \quad \end{bmatrix}}$	Total pxp covariance matrix

(3) In MANOVA we test for group differences by evaluating the ratio of W to T and seeing whether it is significantly less than one, instead of evluating the ratio of the among-group to the within-group variance and seeing whether it is greater than one as we do in the univariate ANOVA. To be specific we evaluate 'Wilk's lambda" which is $\Delta = |W|/|T|$, which is the determinant of W divided by the determinant of T. One test is

$$X^2(p(g-1)df) = -(n-1-p+g/2)\log\Delta .$$

(4) If the null hypothesis H_o = "groups have similar mean vectors" is rejected, and we conclude that the groups are different, then we proceed to do a DA to describe that differences. In matrix algebra the calculations for a DA are simple: we find the roots and vectors of $W^{-1}A$. That is, we invert the matrix W to obtain W^{-1}. Then we do the matrix multiplication to obtain $W^{-1}A$. Then we find the roots and vectors of the $W^{-1}A$ matrix (which is <u>not</u> symmetric). The vectors contain the coefficients in the "discriminant functions" which describe the relationships between the new rotated axes and the original axes, much as the principal component vectors did. However in PCA we were attempting to "most efficiently" describe the

variation and covariation in the data, and to do that we found the roots and vectors of either the covariance or the correlation matrix. In DA we want to most efficiently describe the <u>ratio of among-group to within-group</u> variation and covariation, and to do this we find the roots and vectors of $W^{-1}A$.

8.2 <u>Assignment</u>

To save you the time and bother, here are the matrices and parameters:

$g = 3 \qquad n = 60 \qquad p = 4$

$$W1 = \begin{bmatrix} 796.46 & -987.11 & 190.50 & -51.69 \\ -987.11 & 1244.40 & -257.10 & 62.99 \\ 190.50 & -257.10 & 66.56 & 18.83 \\ -51.69 & 62.99 & -11.29 & 18.83 \end{bmatrix}$$

$$W2 = \begin{bmatrix} 4358.21 & -4413.81 & 55.60 & -284.32 \\ -4413.81 & 4473.19 & -59.38 & 289.04 \\ 55.60 & -59.38 & 3.78 & -4.72 \\ -284.32 & 289.04 & -4.72 & 36.77 \end{bmatrix}$$

$$W3 = \begin{bmatrix} 6.49 & -6.58 & 0.085 & -0.97 \\ -6.58 & 6.67 & -0.099 & 1.20 \\ 0.085 & -0.099 & 0.014 & -0.23 \\ -0.97 & 1.20 & -0.23 & 37.82 \end{bmatrix}$$

$$T = \begin{bmatrix} 57473.6 & -58069.2 & 594.53 & -3584.21 \\ -58069.2 & 58746.3 & -675.90 & 3614.69 \\ 594.53 & -675.90 & 81.35 & -30.40 \\ -3584.21 & 3614.69 & -30.40 & 301.42 \end{bmatrix}$$

8.2.1 Calculate W and A using APL. Then calculate Δ (use PDET to find the determinants) and then the X^2 and the degrees of freedom. Calculate W^{-1} and the $W^{-1}A$. Then use GEIG to find at least the first two roots, and the associated vectors, of $W^{-1}A$.

8.2.2 <u>Now do it all in SAS</u>:

```
TITLE  MANOVA AND DA ON SEDIMENT DATA;
DATA SEDABC;
INPUT PERSAND PERSLTCL PERGRAV PERORG LOCATION;
CARDS;

(the SEDABC data go here - use the 'GET SEDABC DATA' command)

PROC GLM; CLASS LOCATION;
MODEL PERSAND PERSLTCL PERGRAV PERORG=LOCATION;
MANOVA H=LOCATION/PRINTH PRINTE;
PROC CANDISC OUT=DISC; CLASS LOCATION;
VAR PERSAND PERSLTCL PERGRAV PERORG
PROC PLOT; PLOT CAN2*CAN1=LOCATION;
```

8.2.3 Try to interpret these results. Compare your APL results with the SAS results. Also compare this MANOVA/DFA analysis with the PCA analysis.

8.3. Job Listings & Outputs.

```
FILE: MANOVA   SAS      A1  VM/SP - CONVERSATIONAL MONITOR SYSTEM                                        PAGE 001


TITLE SAS ANALYSIS ON SEDIMENT DATA;
DATA SEDABC;
INPUT PERSAND PERSLTCL PERGRAV PERORG LOCATION;
CARDS;
   2.850  97.150   0.0     9.325 1.
   4.622  94.250   1.124   3.477 1.
   3.564  95.944   0.492   8.840 1.
   3.117  96.692   0.191   8.753 1.
   1.513  94.170   4.317   9.451 1.
   2.520  97.480   0.0     6.379 1.
   2.394  96.320   0.235   7.058 1.
   2.460  97.467   0.073   6.894 1.
  31.548  60.702   7.750   6.190 1.
   5.695  93.841   0.464   7.330 1.
   2.865  96.963   0.172   8.734 1.
   8.581  90.078   1.341   6.192 1.
   4.237  95.593   0.170   7.303 1.
   4.329  95.671   0.0     7.560 1.
   2.098  97.344   0.558   7.034 1.
   1.761  98.123   0.165   7.397 1.
   2.309  97.691   0.0     8.377 1.
   4.386  94.620   0.994   7.530 1.
   3.295  95.477   1.228   5.972 1.
   2.953  97.047   0.0     8.591 1.
  46.607  53.003   0.390   5.238 2.
  74.874  24.593   0.533   2.956 2.
  80.096  19.211   0.693   2.412 2.
  81.449  17.401   1.150   2.236 2.
  78.150  20.690   1.160   2.589 2.
  49.475  49.398   1.127   5.073 2.
  47.414  52.418   0.168   5.709 2.
  86.558  12.607   0.835   4.360 2.
  50.699  48.619   0.682   2.275 2.
  80.389  18.145   1.466   2.259 2.
  58.262  41.068   0.670   3.592 2.
  60.382  38.412   1.206   3.933 2.
  55.870  43.314   0.816   2.243 2.
  74.645  23.567   1.788   2.638 2.
  74.689  23.581   1.730   2.735 2.
  42.257  57.207   0.536   5.542 2.
  41.967  57.484   0.549   5.299 2.
  72.371  27.043   0.586   2.055 2.
  76.056  23.602   0.342   2.663 2.
  80.534  18.572   0.894   2.594 2.
   0.650  99.350   0.0     7.908 3.
   0.600  99.400   0.0     7.864 3.
   0.550  99.450   0.0     5.337 3.
   0.985  99.015   0.0     6.975 3.
   0.704  99.296   0.0     7.228 3.
   1.208  98.792   0.0     4.300 3.
   1.469  98.531   0.0     7.524 3.
   0.855  99.145   0.0     7.914 3.
   2.193  97.807   0.0     7.535 3.
   1.832  98.168   0.0     7.298 3.
   0.960  99.005   0.035   6.396 3.
```

```
    0.754  99.188   0.058    6.333 3.
    1.150  98.850   0.0     11.762 3.
    2.078  97.922   0.0      7.526 3.
    1.653  98.347   0.0      7.490 3.
    0.844  99.126   0.030    6.063 3.
    1.668  98.332   0.0      6.086 3.
    1.801  98.178   0.021    6.997 3.
    1.087  98.913   0.0      6.629 3.
    2.504  97.391   0.105    5.854 3.
PROC GLM; CLASS LOCATION;
    MODEL PERSAND PERSLTCL PERGRAV PERORG=LOCATION;
    MANOVA H=LOCATION/PRINTH PRINTE;
PROC CANDISC OUT=DISC; CLASS LOCATION;
    VAR PERSAND PERSLTCL PERGRAV PERORG;
PROC PLOT; PLOT CAN2*CAN1=LOCATION;
```

```
                                                 SAS ANALYSIS ON SEDIMENT DATA                       13:57 THURSDAY, MAY 9, 1985    2
                                                GENERAL LINEAR MODELS PROCEDURE

DEPENDENT VARIABLE: PERSAND

SOURCE                    DF          SUM OF SQUARES          MEAN SQUARE          F VALUE           PR > F         R-SQUARE            C.V.
MODEL                      2          52312.42925440       26156.21462720           288.87           0.0001         0.910199         39.7622
ERROR                     57           5161.15760285          90.54662461                          ROOT MSE                     PERSAND MEAN
CORRECTED TOTAL           59          57473.58685725                                             9.51559901                      23.93125000

SOURCE                    DF               TYPE I SS      F VALUE      PR > F           DF           TYPE III SS      F VALUE      PR > F
LOCATION                   2          52312.42925440       288.87      0.0001            2        52312.42925440       288.87      0.0001
```

```
                                                   SAS ANALYSIS ON SEDIMENT DATA                     13:57 THURSDAY, MAY 9, 1985    3
                                                  GENERAL LINEAR MODELS PROCEDURE

DEPENDENT VARIABLE: PERSLTCL

SOURCE                    DF        SUM OF SQUARES          MEAN SQUARE        F VALUE          PR > F        R-SQUARE            C.V.
MODEL                      2        53021.98620823       26510.99310412         263.99          0.0001        0.902559         13.2812
ERROR                     57         5724.26648170         100.42572775                       ROOT MSE                   PERSLTCL MEAN
CORRECTED TOTAL           59        58746.25268993                                         10.02126378                     75.45463333

SOURCE                    DF             TYPE I SS      F VALUE      PR > F         DF            TYPE III SS      F VALUE      PR > F
LOCATION                   2        53021.99620823       263.99      0.0001          2         53021.98620823       263.99      0.0001
```

```
                                                   SAS ANALYSIS ON SEDIMENT DATA                        13:57 THURSDAY, MAY 9, 1985    4
                                                  GENERAL LINEAR MODELS PROCEDURE

DEPENDENT VARIABLE: PERGRAV

SOURCE                    DF          SUM OF SQUARES           MEAN SQUARE         F VALUE             PR > F        R-SQUARE           C.V.
MODEL                      2             10.99139110            5.49569555            4.45             0.0160        0.135115       180.6659
ERROR                     57             70.35688975            1.23433140                           ROOT MSE                   PERGRAV MEAN
CORRECTED TOTAL           59             81.34828085                                                1.11100468                     0.61495000

SOURCE                    DF               TYPE I SS       F VALUE       PR > F           DF            TYPE III SS       F VALUE       PR > F
LOCATION                   2             10.99139110          4.45       0.0160            2            10.99139110          4.45       0.0160
```

```
                                             SAS ANALYSIS ON SEDIMENT DATA                         13:57 THURSDAY, MAY 9, 1985     5
                                            GENERAL LINEAR MODELS PROCEDURE

DEPENDENT VARIABLE: PERORG

SOURCE                  DF          SUM OF SQUARES          MEAN SQUARE          F VALUE           PR > F          R-SQUARE            C.V.
MODEL                    2            208.00093463         104.00046732            63.46           0.0001          0.690074         21.0559
ERROR                   57             93.41757835           1.63890483                             ROOT MSE                       PERCAG MEAN
CORRECTED TOTAL         59            301.41851298                                               1.28019720                        6.07993333

SOURCE                  DF               TYPE I SS      F VALUE      PR > F          DF             TYPE III SS        F VALUE      PR > F
LOCATION                 2            208.00093463        63.46      0.0001           2            208.00093463          63.46      0.0001
```

```
                                    E = ERROR SSCP MATRIX

DF=57               PERSAND              PERSLTCL               PERGRAV                PERORG

PERSAND        5161.15760235       -5407.49238490          246.17888955         -336.97633555
PERSLTCL      -5407.49238490        5724.26648170         -316.57580430          353.23004425
PERGRAV         246.17888955        -316.57580430           70.35698975          -16.24259370
PERORG         -336.97633555         353.23004425          -16.24259370           93.41757835

        PARTIAL CORRELATION COEFFICIENTS FROM THE ERROR SSCP MATRIX  /  PROB > |R|

                     DF=56          PERSAND  PERSLTCL   PERGRAV    PERORG

                     PERSAND       1.000000 -0.994351  0.408530 -0.495301
                                     0.0000    0.0001    0.0015    0.0001

                     PERSLTCL     -0.994351  1.000000 -0.499844  0.483040
                                     0.0001    0.0000    0.0001    0.0001

                     PERGRAV       0.408530 -0.499844  1.000000 -0.200349
                                     0.0015    0.0001    0.0000    0.1316

                     PERORG       -0.495301  0.483040 -0.200349  1.000000
                                     0.0001    0.0001    0.1316    0.0000
```

```
                         SAS ANALYSIS ON SEDIMENT DATA                    13:57 THURSDAY, MAY 9, 1985    7
                        GENERAL LINEAR MODELS PROCEDURE

                     H = TYPE III SSCP MATRIX FOR: LOCATION

      DF=2                   PERSAND            PERSLTCL             PERGRAV              PERORG

      PERSAND         52312.42025440     -52661.72757160        348.34769720      -3247.23085120
      PERSLTCL       -52661.72757160      53021.79520823       -359.32151080       3261.46488833
      PERGRAV           348.34769720       -359.32151080         10.99139110        -14.15430135
      PERORG          -3247.23085120       3261.46488833        -14.15430135        208.00093463

CHARACTERISTIC ROOTS AND VECTORS OF: E INVERSE * H, WHERE  H = TYPE III SSCP MATRIX FOR: LOCATION     E = ERROR SSCP MATRIX

             CHARACTERISTIC   PERCENT      CHARACTERISTIC VECTOR   V'EV=1
                  ROOT
                                             PERSAND        PERSLTCL         PERGRAV          PERORG

               11.56690295     97.82     -0.34202934     -0.35741647     -0.40504314      0.00248915

                0.26057483      2.18     -0.01861179     -0.01460641     -0.10503852     -0.07043566

                0.00000000      0.00     18.99276148     18.99411355     18.95771705     -0.03018553

                0.00000000      0.00      0.05834420      0.05223041     -0.02554548      0.09013353
```

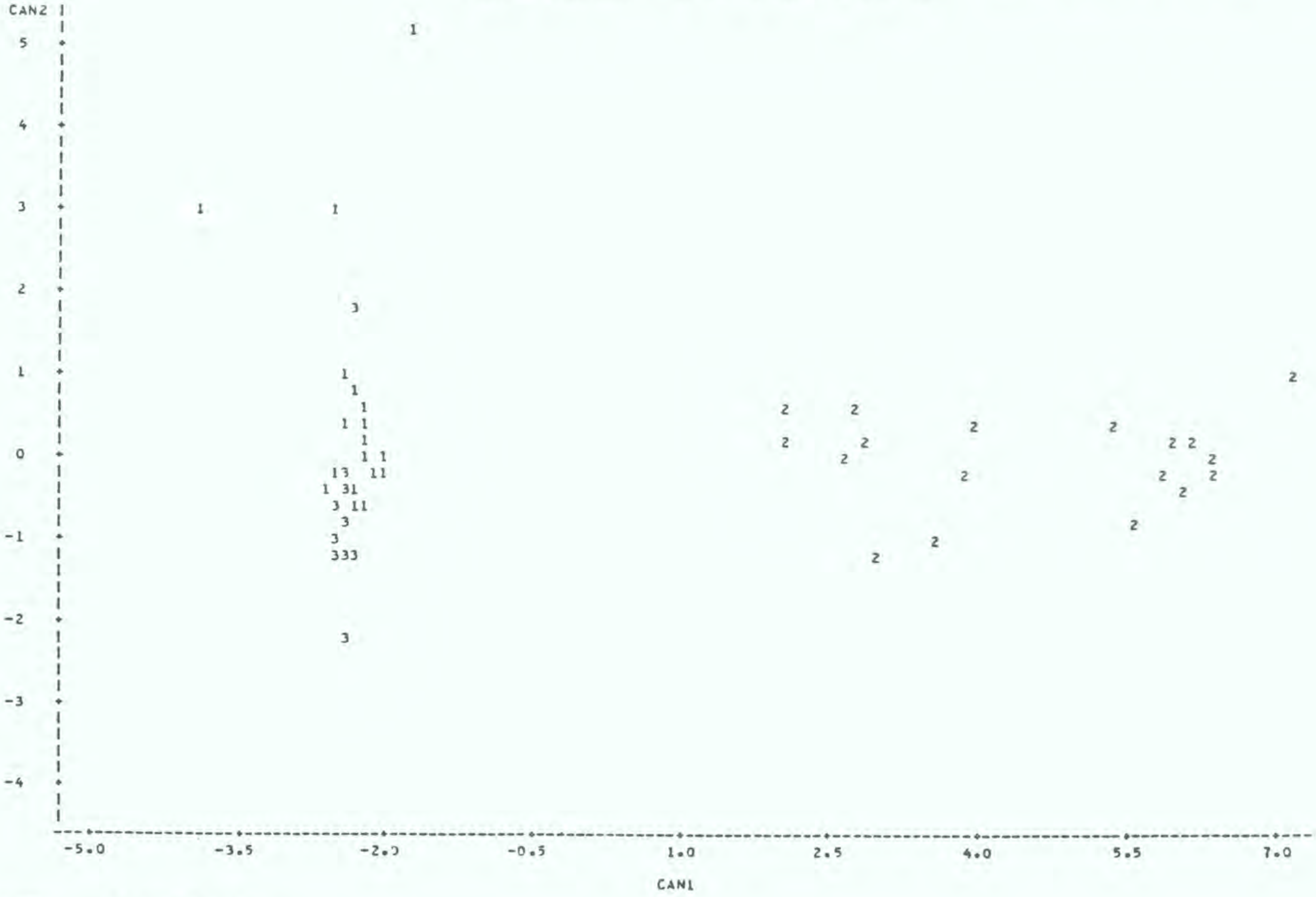
```
                          SAS ANALYSIS ON SEDIMENT DATA                     13:57 THURSDAY, MAY 9, 1985    8
                         GENERAL LINEAR MODELS PROCEDURE

         MANOVA TEST CRITERIA FOR THE HYPOTHESIS OF NO OVERALL LOCATION EFFECT

                                H = TYPE III SSCP MATRIX FOR: LOCATION
                                E = ERROR SSCP MATRIX
                                P = DEP. VARIABLES =        4
                                Q = HYPOTHESIS DF  =        2
                                NE= DF OF E        =       57
                                S = MIN(P,Q)       =        2
                                M = .5(ABS(P-Q)-1) =      0.5
                                N = .5(NE-P-1)     =     26.0

-----------------------------------------------------------------------------------------------------

HOTELLING-LAWLEY TRACE = TR(E**-1*H) =            11.92747793     (SEE PILLAI'S TABLE #3)

        F APPROXIMATION = 2(S*N+1)*TR(E**-1*H)/(S*S*(2M+S+1))    WITH S(2M+S+1) AND 2(S*N+1) DF

              F(8,106) =     79.02      PROB > F = 0.0001

-----------------------------------------------------------------------------------------------------

PILLAI'S TRACE         V = TR(H*INV(H+E)) =        1.12776525     (SEE PILLAI'S TABLE #2)

        F APPROXIMATION = (2N+S+1)/(2M+S+1) * V/(S-V)            WITH S(2M+S+1) AND S(2N+S+1) DF

              F(8,110) =     17.78      PROB > F = 0.0001

-----------------------------------------------------------------------------------------------------

WILKS' CRITERION       L = DET(E)/DET(H+E) =       0.06252690     (SEE RAO 1973 P 555)

        EXACT F = (1-SQRT(L))/SQRT(L)*(NE+Q-P-1)/P               WITH 2P AND 2(NE+Q-P-1) DF

              F(8,108) =     40.45      PROB > F = 0.0001

-----------------------------------------------------------------------------------------------------

ROY'S MAXIMUM ROOT CRITERION =                    11.56693295     (SEE AMS VOL 31 P 625)

        FIRST CANONICAL VARIABLE YIELDS AN F UPPER BOUND

              F(2,57) =    332.51      (UPPER BOUND)

-----------------------------------------------------------------------------------------------------
```

```
                          SAS ANALYSIS ON SEDIMENT DATA                         13:57 THURSDAY, MAY 9, 1985    5
                          CANONICAL DISCRIMINANT ANALYSIS

                    60 OBSERVATIONS          59 DF TOTAL
                     4 VARIABLES             57 DF WITHIN CLASSES
                     3 CLASSES                2 DF BETWEEN CLASSES

CANONICAL CORRELATIONS AND TESTS OF HO: THE CANONICAL CORRELATION IN THE CURRENT ROW AND ALL THAT FOLLOW ARE ZERO

    CANONICAL     ADJUSTED     APPROX       VARIANCE   CANONICAL    LIKELIHOOD
   CORRELATION    CAN CORR    STD ERROR       RATIO    R-SQUARED      RATIO      F STATISTIC  NUM DF   DEN DF  PROB>F

1  0.959715637  0.954743941  0.010277380    11.6567  0.921054104  0.062626899      40.4453        8      108  0.0000
2  0.454654973  0.369637709  0.103277412     0.2606  0.206711149  0.793288951       4.7772        3       55  0.0050


                      MULTIVARIATE TEST STATISTICS AND F APPROXIMATIONS

STATISTIC                     VALUE              F            NUM DF            DEN DF             PROB>F

WILKS' LAMBDA              0.0626269       40.44525              8               108        4.08791E-29
PILLAI'S TRACE             1.127765        17.77321              8               110        9.69357E-17
HOTELLING-LAWLEY TRACE    11.92748         79.01954              8               106        3.75164E-41
ROY'S GREATEST ROOT       11.6669         160.4199               4                55        1.25025E-29

                    NOTE: F STATISTIC FOR ROY'S GREATEST ROOT IS AN UPPER BOUND
                          F STATISTIC FOR WILKS' LAMBDA IS EXACT


                             TOTAL CANONICAL STRUCTURE

                                        CAN1          CAN2

                       PERSAND        0.9933        0.0841
                       PERSLTCL      -0.9885       -0.1103
                       PERGRAV        0.1622        0.7324
                       PERORG        -0.8575        0.2490


                            BETWEEN CANONICAL STRUCTURE

                                        CAN1          CAN2

                       PERSAND        0.9992        0.0401
                       PERSLTCL      -0.9986       -0.0528
                       PERGRAV        0.4234        0.9059
                       PERORG        -0.9907        0.1363


                             WITHIN CANONICAL STRUCTURE

                                        CAN1          CAN2

                       PERSAND        0.9313        0.2499
                       PERSLTCL      -0.8398       -0.3148
                       PERGRAV        0.0490        0.7015
                       PERORG        -0.4328        0.3984
```

```
                SAS ANALYSIS ON SEDIMENT DATA
               CANONICAL DISCRIMINANT ANALYSIS

            STANDARDIZED CANONICAL COEFFICIENTS

                             CAN1            CAN2

            PERSAND       -80.5952         1839.49
            PERSLTCL      -85.1493         1909.34
            PERGRAV        -3.5908         71.8523
            PERORG          0.0425          1.2020

                  RAW CANONICAL COEFFICIENTS

                             CAN1                   CAN2

     PERSAND          -2.592265143           60.539147745
     PERSLTCL         -2.598435433           60.508907759
     PERGRAV          -3.058008384           61.191655247
     PERORG            0.013792664            0.531777736

             CLASS MEANS ON CANONICAL VARIABLES

             LOCATION             CAN1             CAN2

                    1          -2.3143           0.5128
                    2           4.7080          -0.0069
                    3          -2.3937          -0.6059
```

SAS ANALYSIS ON SEDIMENT DATA 13:57 THURSDAY, MAY 9, 1985 11

PLOT OF CAN2*CAN1 SYMBOL IS VALUE OF LOCATION



NOTE: 14 OBS HIDDEN

9. PRINCIPLES OF SAMPLING DESIGN

9.1 Ten Principles

1. Be able to state concisely to someone else what question you are asking. Your results will be as coherent and as comprehensible as your initial conception of the problem.

2. Take replicate samples within each combination of time, location, and any other controlled variable. Differences among can only be demonstrated by comparison to differences within.

3. Take an equal number of randomly allocated replicate samples for each combination of controlled variables. Putting samples in "representative" or "typical" places is not random sampling.

4. To test whether a condition has an effect, collect samples both where the condition is present and where the condition is absent but all else is the same. An effect can only dbe demonstrated by comparison with a control.

5. Carry out some preliminary sampling to provide a basis for evaluation of sampling design and statistical analysis options. Those who skip this step because they do not have enough time usually end up losing time.

6. Verify that your sampling device or method is sampling the population you think you are sampling, and with equal and adequate efficiency over the entire range of sampling conditions to be encountered. Variation in efficiency of sampling from area to area biases among-area comparisons.

7. If the area to be sampled has a large-scale environmental attern, break the area up into relatively homogeneous subareas and allocate samples to each in

proportion to the size of the subarea. If it is an estimate of totasl abundance over the entire area that is desired, make the allocation proportional to the number of organisms in the subarea.

8. Verify that your sample unit size is appropriate to the size, densities, and spatial distributions of the organisms you are sampling. Then estimate the number of replicate samples required to obtain the precision you want.

9. Test your data to determine whether the error variation is homogeneous, normally distributed, and independent of the mean. If it is not, as will be the case for most field data, then (a) appropriatelyd transform the data, (b) a distribution-free (nonparametric) procedure,(c)use an appropriate sequential sampling design, or (d)test against simulated H_o data.

10. Having chosen the best statistical method to test your hypothesis, stick with the result. An unexpected or undesired result is not a valid reason for rejecting the method and hunting for a "better" one.

9.2 Estimation of sample number

9.2.1 Based on preliminary sampling:
Say that preliminary sampling estimates $\bar{X}_1$=18 and S_1^2=236. If you wish to collect enough samples to estimate $\bar{X}_2$ so that the true mean μ lies within ±20% of X with a chance of α=0.05 or less that it doesn't, then

$$\bar{X} \pm t\ (S.E.) = \bar{X} \pm t\sqrt{S^2}/n = X \pm tS/\sqrt{n}$$

which should equal $\bar{X} \pm 0.2\bar{X}$.

Therefore $t\,S/\sqrt{n} = 0.2\,\bar{X}$ and, if n is fairly large,

$$(2)\ 15.36/\sqrt{n} \approx (0.2)(18) \text{ and } n \approx 73.$$

(The value of $t_{\alpha=0.5}$ for 72 df is almost exactly 2.)

9.2.2 Without preliminary sampling, but assuming Taylor's Power Law:

$S^2 = a\bar{X}^{b}$, with $a \approx 1$ and $b \approx 2$:
Define $D_o = S.E./\bar{X} = \frac{S/\sqrt{n}}{\bar{X}}$.

If $a \approx 1$ and $b \approx 2$, then $S^2 = a\bar{X}^{b} \approx \bar{X}^2$ and $S \approx \bar{X}$.

$$\text{Therefore } D_o = \frac{S/\sqrt{n}}{\bar{X}} \qquad \frac{X/\sqrt{n}}{\bar{X}} = \frac{1}{\sqrt{n}} \text{ and } n \approx \frac{1}{D_o}\,2$$

If we want a precision of ±20% with $\alpha=0.05$,

$$(2)\ (S.E.) \approx 0.2\,\bar{X} \text{ as before, and}$$

$$(2)\ S.E./\bar{X} = 2\,D_o \approx 0.2\ .$$

$$\text{Therefore } D_o \approx 0.1 \text{ and } n \approx \frac{1}{D_o^2} = \frac{1}{(0.1)^2} = 100$$

Bibliography

Anscombe, F.J. 1981. Computing in statistical science through APL. Springer-Verlag, New York. 426 p.

Atchley, W.R., and E.H. Bryant, eds. 1975. Multivariate statistical methods: among-groups covariation. Dowden, Hutchinson & Ross, Stroudsburg, Pennylvania. 464 p.

Conley, W.E. 1982. BASIC for beginners. Petrocelli Books, New York 162 p.

Cooley, W.W., and P.R. Lohnes. 1971. Multivariate data analysis. Wiley, New York. 364 p.

Davis, J.C. 1973. Statistics and data analysis in geology. Wiley, New York.

Draper, N.R., and H. Smith. 1966. Applied regression analysis. Wiley, New York. 407 p.

Elliott, J.M. 1977. Some methods for the statistical analysis of samples of benthic invertebrates. 2nd ed. Freshwater Biological Assoc. U.K., Sci. Publ. No.25. The Ferry House, Ambleside, Cumbria, UK. 156 p.

Gauch, H.G., Jr. 1982. Multivariate analysis in community ecology. Cambridge, U.K. 298 p.

Gilman, L., and A.J. Rose. 1976. APL: an interactive approach. Wiley, New York. 378 p.

Gomez, A.C. 1983. The basics of BASIC. Holt, Rinehart and Winston, New York. 303 p.

Green, P.E., and J.D. Carroll. 1976. Mathematical tools for applied multivariate analysis. Academic, New York. 376 p.

Green, R.H. 1979. Sampling design and statistical methods for environmental biologists. Wiley, New York. 257 p.

Jeffers, J.N.R. 1982. Modelling. Chapman and Hall, London. 80 p.

Lamoitier, J.P. 1981. Fifty BASIC exercises. Sybex, Europe. 231 p.

Lee, J.D. and T.D. Lee. 1982. Statistics and computer methods in BASIC. Von Nostrand Reinhold, New York. 198 p.

Legendre, L., and P. Legendre. 1983. Numerical ecology. Elsevier, Amsterdam. 419 p.

McCracken, D.D. 1974. A simplified guide to FORTRAN programming. Wiley, New York. 278 p.

McNeil, D.R. 1977. Interactive data analysis: a practical primer. Wiley, New York. 186 pp.

Moore, R.W. 1978. Introduction to the use of computer packages for statistical analysis. Prentice-Hall, Englewood Cliffs, New Jersey. 115 p.

Orloci, L. 1978. Multivariate analysis in vegetation research. 2nd ed. Junk, The Hague. 451 p.

Orloci, L., and N.C. Kenkel. 1984. Introduction to data analysis, with applications in population and community biology. Univ. of Western Ontario, London, Canada. (Prepublication edition).

Pommier, S. 1983. An introduction to APL. Cambridge, U.K. 136 p.

Ramsey, J.B., and G.L. Musgrave. 1981. APL-STAT: a do-it-yourself guide to computational statistics using APL. Wadsworth, Belmont, California. 339 p.

Ryan, T.A., B.L. Joiner, and B.F. Ryan. 1976. Minitab student handbook. Duxbury Press, Boston.

Ryan, T.A., B.L. Joiner, and B.F. Ryan. 1976. Minitab Reference Manual. Minitab Project, Pennsylvania State Univ., University Park, Pennsylvania. 138 p.

SAS User's Guide. 1979. SAS Institute, Inc. Raleigh, North Carolina. 494 p.

SAS User's Guide: Statistics. 1982. SAS Institutee, Inc., North Carolina. 584 p.

Slater, L.J. 1971. First steps in basic FORTRAN. Chapman and Hall, London. 104 p.

Snedecor, G.W., and W.G. Cochran. 1980. Statistical methods. Iowa State Univ. Press, Ames, Iowa. 507 p.

Sokal, R.R., and F.J. Rohlf. 1973. Introduction to biostatistics. Freeman, San Francisco. 368 p.

Southwood, T.R.E. 1978. Ecological Methods. Chapman and Hall, London. 524 p.

Spain, J.D. 1982. BASIC microcomputer models in biology. Addison-Wesley, Reading, Massachusetts. 354 p.

Steel, R.G.D., and J.H. Torrie. 1960. Principles and procedures of statistics with special reference to the biological sciences. McGraw-Hill, New York. 481 p.

Velleman, P.F., and D.C. Hoaglin. 1981. Applications, basics, and computing of exploratory data analysis. Duxbury Press, Boston. 354 p.

APPENDIX I. - AGENDA

OPENING SESSION

Addresses by: K. L. Chan, Acting Head, Zoology Department
National Universsity of Singapore

: J. R. E. Harger, Unesco-Mab/Unep representative
from ROSTSEA

: H. H. Huang, Deputy Vice-Chancellor National
University of Singapore

COURSE SCHEDULE

	Morning	Afternoon
April 22 M	Introductory remarks - hand out schedule.	Tour of facilities, illustrate use of equipment, hand out out Tu-W tutorial.
23 Tu	Course organization, objectives, assumptions re. background, review of simple linear regression analysis.	Doing linear regression & plots using the various hardware & software.
24 W	Principles of linear, regression analysis, Models I & II, "Cookbook" versus matrix algebra solutions — latter in MINITAB & APL.	Continuation - if there is time, try it with larger data set, and/or transformation of variables. Explore MINITAB, APL.
25 Th	Introduction to common bivariate relationships in biology - nonlinear models.	Demonstration of doing regression by matrix algebra in MINITAB and APL. Demonstration of MINITAB plot & regression options. Introduction to SAS.

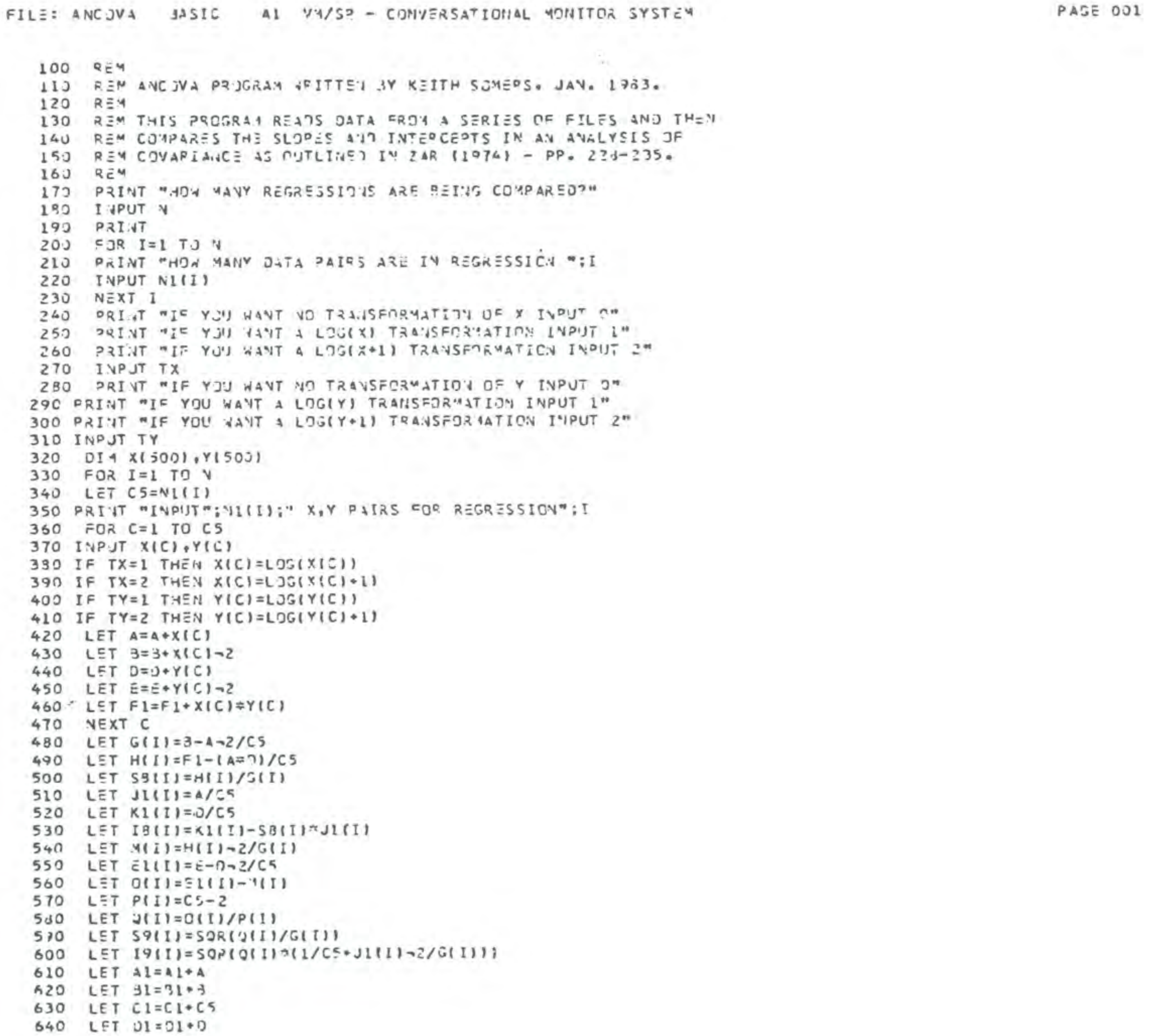
	26	F	Continuation of presentation of biologically important nonlinear bivariate models. Intro. to ratio variables.	Doing linear and nonlinear regression analysis - emphasis on MINITAB, some SAS. Includes matrix algebra approach.
	27	Sa	Unscheduled - please use to your best advantage.	———
	29	M	Ratio variables, Taylor's Power Law (re. choosing transformations), introduction to analysis of covariance.	Doing nonlinear regression models with asymptotes, emphasis on MINITAB.
	30	Tu	Analysis of covariance, Walford plots, transforming multivariate data.	Doing a nonlinear analysis using micros. Doing a ratio variable problem.
May	1	W	Review of material covered so far (please have questions ready!). Overview of statistical models, including multivariate.	Doing a Walford plot problem in MINITAB.
	2	Th	Continue introduction to multivariate models - emphasis on tests for structure, ordination & clustering.	Doing an analysis of covariance in MINITAB, SAS, and on micros.
	3	F	Ordination & clustering, and introduction to MV ANOVA and discriminant analysis.	Calculation of matrices for MV analysis, testing for structure - emphasis on MINITAB.

4	Sa	Unscheduled - please use to your best advantage.	________
6	M	MV ANOVA and discriminant analysis.	Doing MV structure tests, ordination, in MINITAB & SAS.
7	Tu	Canonical correlation analysis Introduction to principles of sampling design.	Doing ordination & clustering-related analyses using custom-written programs, including on micros.
8	W	Continuation on principles of sampling design - sample unit size, estimation of necessary number of samples.	Doing MANOVA and DFA in SAS and APL, and canonical correlation in SAS.
9	Th	Transforming data - rationale, principles, strategy. Begin discussion of examples of design problems.	Doing MANOVA/DFA on APPLES. Doing exercise in sampling random and contagious distributions.
10	F	Review of course. Course evaluation. Discussion of participants' design & biostatistical analysis problems "back home".	Continuation of morning discussion of "back home" case studies. Clean up. Make sure you have disks.

CLOSING SESSION

Addresses by	: Kuswata Kartawinata, Unesco/Mab/Unep representative from ROSTSEA.
	: T. W. Chen, Acting Head, Zoology Department National University of Singapore.
	: S. D. Tandjung, course participant representative.

APPENDIX II -- PROGRAMS



```
FILE: ANCOVA   BASIC    A1  VM/SP - CONVERSATIONAL MONITOR SYSTEM                                    PAGE 001

  100  REM
  110  REM ANCOVA PROGRAM WRITTEN BY KEITH SOMERS. JAN. 1983.
  120  REM
  130  REM THIS PROGRAM READS DATA FROM A SERIES OF FILES AND THEN
  140  REM COMPARES THE SLOPES AND INTERCEPTS IN AN ANALYSIS OF
  150  REM COVARIANCE AS OUTLINED IN ZAR (1974) - PP. 228-235.
  160  REM
  170  PRINT "HOW MANY REGRESSIONS ARE BEING COMPARED?"
  180  INPUT N
  190  PRINT
  200  FOR I=1 TO N
  210  PRINT "HOW MANY DATA PAIRS ARE IN REGRESSION ";I
  220  INPUT N1(I)
  230  NEXT I
  240  PRINT "IF YOU WANT NO TRANSFORMATION OF X INPUT 0"
  250  PRINT "IF YOU WANT A LOG(X) TRANSFORMATION INPUT 1"
  260  PRINT "IF YOU WANT A LOG(X+1) TRANSFORMATION INPUT 2"
  270  INPUT TX
  280  PRINT "IF YOU WANT NO TRANSFORMATION OF Y INPUT 0"
 290 PRINT "IF YOU WANT A LOG(Y) TRANSFORMATION INPUT 1"
 300 PRINT "IF YOU WANT A LOG(Y+1) TRANSFORMATION INPUT 2"
 310 INPUT TY
 320  DIM X(500),Y(500)
 330  FOR I=1 TO N
 340  LET C5=N1(I)
 350 PRINT "INPUT";N1(I);" X,Y PAIRS FOR REGRESSION";I
 360  FOR C=1 TO C5
 370 INPUT X(C),Y(C)
 380 IF TX=1 THEN X(C)=LOG(X(C))
 390 IF TX=2 THEN X(C)=LOG(X(C)+1)
 400 IF TY=1 THEN Y(C)=LOG(Y(C))
 410 IF TY=2 THEN Y(C)=LOG(Y(C)+1)
 420  LET A=A+X(C)
 430  LET B=B+X(C)¬2
 440  LET D=D+Y(C)
 450  LET E=E+Y(C)¬2
 460  LET F1=F1+X(C)*Y(C)
 470  NEXT C
 480  LET G(I)=B-A¬2/C5
 490  LET H(I)=F1-(A*D)/C5
 500  LET S8(I)=H(I)/G(I)
 510  LET J1(I)=A/C5
 520  LET K1(I)=D/C5
 530  LET I8(I)=K1(I)-S8(I)*J1(I)
 540  LET M(I)=H(I)¬2/G(I)
 550  LET E1(I)=E-D¬2/C5
 560  LET O(I)=E1(I)-M(I)
 570  LET P(I)=C5-2
 580  LET Q(I)=O(I)/P(I)
 590  LET S9(I)=SQR(Q(I)/G(I))
 600  LET I9(I)=SQR(Q(I)*(1/C5+J1(I)¬2/G(I)))
 610  LET A1=A1+A
 620  LET B1=B1+B
 630  LET C1=C1+C5
 640  LET D1=D1+D
 650  LET E2=E2+E
 660  LET F2=F2+F1
 670  LET A3=A3+G(I)
 680  LET B3=B3+H(I)
 690  LET C3=C3+E1(I)
 700  LET S4=S4+O(I)
 710  LET F4=F4+P(I)
 720 LET A=0
 730 LET B=0
 740 LET D=0
 750 LET E=0
 760 LET F1=0
 770  NEXT I
 780  LET S3=C3-B3¬2/A3
 790  LET F3=C1-N-1
 800  LET A6=B1-A1¬2/C1
 810  LET B6=F2-(A1*D1)/C1
 820  LET C6=E2-D1¬2/C1
 830  LET S1=C6-B6¬2/A6
 840  LET F2=C1-2
 850  LET R1=B6/A6
 860  LET R2=(D1/C1)-(R1*(A1/C1))
 870  LET M7=S4/F4
 880  LET M3=S3/F3
 890  LET F8=((S3-S4)/(N-1))/M7
 900  LET F9=((S1-S3)/(N-1))/M3
 910  LET F5=N-1
 920  PRINT
 930  IF N<>2 THEN 1170
 940  PRINT "THE ANALYSIS OF COVARIANCE BETWEEN"
 950  PRINT "REGRESSION ONE AND REGRESSION TWO"
 960  PRINT "HAS PRODUCED THE FOLLOWING: "
 970  LET J2=S4/F4
 980  LET J3=SQR(J2/G(1)+J2/G(2))
 990  LET J4=ABS(S8(1)-S8(2))/J3
 1000  LET J8=B3/A3
 1010  LET J5=ABS(K1(1)-K1(2))-J8*(ABS(J1(1)-J1(2)))
 1020  LET J6=SQR((S3/F3)*(1/N1(1)+1/N1(2)+(J1(1)-J1(2))¬2/A3))
 1030  LET J7=J5/J6
 1040  PRINT
 1050  PRINT "THE STUDENT'S T STATISTIC FOR B1=B2 WITH A "
 1060  PRINT "TWO-TAILED ALPHA OF 0.05 AND  ";F4;"  DEGREES OF FREEDOM"
 1070  PRINT "IS:  ";J4
 1080  PRINT
 1090  PRINT "IF B1=B2 IS FALSE, THEN TWO DIFFERENT POPULATIONS WERE SAMPLED."
 1100  PRINT "IF B1=B2, THEN TEST FOR COMMON INTERCEPTS."
 1110  PRINT
 1120  PRINT "THE STUDENT'S T STATISTIC FOR COMMON INTERCEPTS WITH A "
 1130  PRINT "TWO-TAILED ALPHA OF 0.05 AND  ";F3;"  DEGREES OF FREEDOM"
 1140  PRINT "IS:  ";J7
 1150  PRINT
 1160  GO TO 1280
 1170  PRINT "THE ANALYSIS OF COVARIANCE HAS BEEN COMPLETED"
 1180  PRINT "FOR THE ";N;" REGRESSION LINES."
 1190  PRINT
 1200  PRINT "THE ANCOVA TEST AS OUTLINED IN ZAR (1974) CHAPTER 17,"
 1210  PRINT "PRODUCES TWO F-STATISTICS THAT DETERMINE IF THE COMBINED"
 1220  PRINT "SLOPES AND COMBINED INTERCEPTS ARE DIFFERENT."
 1230  PRINT
 1240  PRINT "THE F-VALUE FOR THE SLOPES IS: ";F8
 1250  PRINT "THE F-VALUE FOR THE INTERCEPTS IS: ";F9
 1260  PRINT "BOTH VALUES HAVE A NUMERATOR D. OF F. OF: ";F5
 1270  PRINT "AND A DENOMINATOR D. OF F. OF: ";F3
 1280  PRINT
 1290  PRINT "IF THE SLOPES ARE NOT SIGNIFICANTLY DIFFERENT, "
 1300  PRINT "THE COMMON REGRESSION SLOPE IS: ";R1
 1310  PRINT
 1320  PRINT "AND IF THE INTERCEPTS ARE NOT SIGNIFICANTLY DIFFERENT, "
 1330  PRINT "THE COMMON REGRESSION INTERCEPT IS: ";R2
 1340  IF N=2 THEN 2030
 1350  PRINT
 1360  PRINT "IF THE SLOPES OR INTERCEPTS ARE DIFFERENT, THEN "
 1370  PRINT "A MULTIPLE RANGE TEST CAN BE USED TO IDENTIFY THE "
 1380  PRINT "DIFFERENCES BETWEEN THE SET OF REGRESSIONS. "
 1390  PRINT
 1400  PRINT "DO YOU WANT TO DO THE MULTIPLE RANGE TESTS?"
 1410  PRINT "TYPE Y OR N. "
 1420  INPUT A9$
 1430  IF A9$="N" THEN 2030
 1440  REM TO COMPLETE THE MULTIPLE RANGE TESTS, SEVERAL PROCEDURES
 1450  REM ARE AVAILABLE.  TYPICALLY THE NEWMAN-KEULS MULTIPLE RANGE
 1460  REM TEST IS USED IF EACH REGRESSION IS COMPARED WITH EACH
 1470  REM OTHER REGRESSION (OPTION 1).  HOWEVER, IF THE REGRESSIONS
 1480  REM ARE BASED ON DIFFERENT X-VALUES THEN A DIFFERENT FORMULA
 1490  REM MUST BE USED (OPTION 2).  ALTERNATIVELY, IF ONE OF THE
 1500  REM REGRESSION LINES IS A CONTROL AND ALL OTHERS ARE TO BE
 1510  REM COMPARED TO IT, DUNNETT'S TEST IS APPROPRIATE (OPTION 3).
 1520  REM AGAIN, IF THE X-VALUES ARE DIFFERENT, AN ALTERNATIVE
 1530  REM FORMULA IS REQUIRED (OPTION 4).
 1540  PRINT "CHOOSE A MULTIPLE RANGE TEST FROM THIS LIST: "
 1550  PRINT "(1) NEWMAN-KEULS WITH THE SAME X-VALUES"
 1560  PRINT "(2) NEWMAN-KEULS WITH DIFFERENT X-VALUES"
 1570  PRINT "(3) DUNNETT'S TEST WITH THE SAME X-VALUES"
 1580  PRINT "(4) DUNNETT'S TEST WITH DIFFERENT X-VALUES"
 1590  INPUT A8
 1600  PRINT
 1610  PRINT
 1620  PRINT "THE RESULTS OF THE MUMLTIPLE RANGE TESTS"
 1630  PRINT "ARE AS FOLLOWS: "
 1640  PRINT
 1650  REM TO ACCURATELY DEFINE THE Q-STATISTIC GIVEN BELOW,
 1660  REM YOU MUST RANK THE SLOPES AND INTERCEPTS FOR EACH
 1670  REM REGRESSION.  THE DIFFERENCE IN ORDER BETWEEN PAIRS
 1680  REM OF REGRESSIONS PROVIDES A P-VALUE FOR D. OF F.
 1690  REM NEEDED" IN THE Q-TABLE (I.E. Q(0.05)(DF;V)(DF;P)).
 1700  REM IF HI-TO-LOW RANKING SEPARATES REGS. 1 AND 4 BY 3 VALUES,
 1710  REM THEN P=5 (I.E. FOR 135246, REGS. 1 + 4 HAVE P=5).
 1720  PRINT " REGRESSION    SLOPE     ELEVATION-(INT)"
 1730  PRINT
 1740  FOR I=1 TO N
 1750  PRINT "    ";I;"        ";S1(I);"    ";I8(I)
 1760  NEXT I
 1770  PRINT
 1780  PRINT " REGRESSIONS      SLOPE-Q      ELEVATION-Q  D.F.(V)"
 1790  PRINT
 1800  FOR I=1 TO (N-1)
 1810  LET K=I+1
 1820  FOR J=K TO N
 1830  LET B8=(H(I)+H(J))/(G(I)+G(J))
 1840  IF A3>1 THEN 1860
 1850  LET X8=SQR(M7/G(I))
 1860  IF A3<>2 THEN 1880
 1870  LET X8=SQR((M7/2)*(1/G(I)+1/G(J)))
 1880  IF A3<>3 THEN 1900
 1890  LET X8=SQR(2*M7/G(I))
 1900  IF A3<>4 THEN 1920
 1910  LET X8=SQR(M7*(1/G(I)+1/G(J)))
 1920  LET Q8=ABS(S8(I)-S8(J))/X8
 1930  IF A3=1 THEN 1950
 1940  IF A3<>2 THEN 1960
 1950  LET Y8=SQR(M3/2*(1/N1(I)+1/N1(J)+(J1(I)-J1(J))¬2/(G(I)+G(J))))
 1960  IF A3<3 THEN 1990
 1970  IF A3>4 THEN 1990
 1980  LET Y8=SQR(M3*(1/N1(I)+1/N1(J)+(J1(I)-J1(J))¬2/(G(I)+G(J))))
 1990  LET R8=ABS((K1(I)-K1(J))-B8*(J1(I)-J1(J)))/Y8
 2000  PRINT "  ";I;" AND ";J;"    ";Q8;"    ";R8;"    ";F4
 2010  NEXT J
 2020  NEXT I
 2030  END
```

```
   100  REM
   110  REM LINEAR REGRESSION PROGRAM WRITTEN BY KEITH SOMERS, MAY 1982.
   120  REM
   130  REM SIMPLE LINEAR REGRESSION PROGRAM THAT RECEIVES DATA FROM THE
   140  REM KEYBOARD AND COMPUTES ADVANCED STATISTICS FOR THAT DATA.
   150  REM
   160  REM THE STATISTICS AND REGRESSION ANALYSIS FOLLOW THE CHAPTER
   170  REM ON THAT SUBJECT IN "BIOSTATISTICAL ANALYSIS" BY ZAR
   180  REM
  190  DIM X(200),Y(200)
  200  DIM Y1(200),Y2(200)
  210  PRINT "HOW MANY X-Y PAIRS DO YOU WANT TO ENTER?"
  220  INPUT N
  230  PRINT
   240  PRINT "IF YOU WANT NO TRANSFORMATION OF X INPUT 0"
   250  PRINT "IF YOU WANT A LOG(X) TRANSFORMATION INPUT 1"
   260  PRINT "IF YOU WANT A LOG(X+1) TRANSFORMATION INPUT 2"
   270  INPUT TX
   280  PRINT "IF YOU WANT NO TRANSFORMATION OF Y INPUT 0"
  290 PRINT "IF YOU WANT A LOG(Y) TRANSFORMATION INPUT 1"
  300 PRINT "IF YOU WANT A LOG(Y+1) TRANSFORMATION INPUT 2"
  310 INPUT TY
  320  PRINT
  330  PRINT "ENTER THE DATA AS X-Y PAIRS."
  340  PRINT
  350  FOR C=1 TO N
  360  INPUT X(C),Y(C)
  370 IF TX=1 THEN X(C)=LOG(X(C))
  380 IF TX=2 THEN X(C)=LOG(X(C)+1)
  390 IF TY=1 THEN Y(C)=LOG(Y(C))
  400 IF TY=2 THEN Y(C)=LOG(Y(C)+1)
  410  NEXT C
  420  LET Z5=N-2
  430  PRINT
  440  PRINT "WHAT IS THE T-VALUE FOR THE 95% CONFIDENCE LIMITS"
  450  PRINT "WITH  ";Z5;"  DEGREES OF FREEDOM?"
  460  INPUT T1
  470  PRINT
  480  PRINT
  490  FOR C=1 TO N
  500  REM A IS THE SUM OF THE X VALUES
  510  LET A=A+X(C)
  520  REM B IS THE SUM OF SQUARED X VALUES
  530  REM ZAR'S BIG X-SQUARED
  540  LET B=B+X(C)¬2
  550  REM D IS THE SUM OF THE Y VALUES
  560  LET D=D+Y(C)
  570  REM E IS THE SUM OF SQUARED Y VALUES
  580  REM ZAR'S BIG Y-SQUARED
  590  LET E=E+Y(C)¬2
  600  REM F IS THE SUM OF X*Y
  610  REM ZAR'S BIG XY
  620  LET F=F+X(C)*Y(C)
  630  NEXT C
  640  REM G IS THE SUM OF SQUARES OF X
```

```
 650  REM ZAR'S LITTLE X-SQUARED
 660  LET G=B-A¬2/N
 670  REM H IS THE SUM OF CROSS-PRODUCT DEVIATIONS
 680  REM ZAR'S LITTLE XY
 690  LET H=F-(A*D)/N
 700  REM I IS THE SLOPE
 710  LET I=H/G
 720  REM J IS THE MEAN OF X
 730  LET J=A/N
 740  REM K IS THE MEAN OF Y
 750  LET K=D/N
 760  REM L IS THE Y-INTERCEPT
 770  LET L=K-I*J
 780  REM M IS THE SUM OF SQUARES OF THE REGRESSION
 790  REM M IS ALSO THE REGRESSION MEAN SQUARE
 800  LET M=H¬2/G
 810  REM E1 IS THE TOTAL SUM OF SQUARES
 820  REM ZAR'S LITTLE Y-SQUARED
 830  LET E1=E-D¬2/N
 840  REM N1 IS R-SQUARED, THE COEFFICIENT OF DETERMINATION
 850  LET N1=M/E1
 860  REM N2 IS R, THE CORRELATION COEFFICIENT
 870  LET N2=SQR(N1)
 880  REM O IS THE RESIDUAL SUM OF SQUARES
 890  LET O=E1-M
 900  REM P IS THE RESIDUAL DEGREES OF FREEDOM
 910  LET P=N-2
 920  REM Q IS THE RESIDUAL MEAN SQUARE
 930  LET Q=O/P
 940  REM R IS THE F-STATISTIC TO DETERMINE IF THE SLOPE EQUALS ZERO
 950  LET R=M/Q
 960  REM S IS THE STANDARD ERROR OF THE ESTIMATE (EPSILON)
 970  LET S=SQR(Q)
 980  REM I1 IS THE STANDARD ERROR OF THE SLOPE WHICH IS USED TO TEST
 990  REM FOR SIGNIFICANCE OF THE SLOPE AS RELATED TO A SPECIFIED VALUE
 1000  LET I1=SQR(Q/G)
 1010  REM I2 AND I3 ARE 95% CONFIDENCE LIMITS AROUND THE SLOPE
 1020  LET I2=I-T1*I1
 1030  LET I3=I+T1*I1
 1040  REM L1 IS THE STANDARD ERROR OF THE INTERCEPT
 1050  LET L1=SQR(Q*(1/N+J¬2/G))
1060  REM L2 AND L3 ARE THE 95% C.L. FOR THE INTERCEPT
1070  LET L2=L-T1*L1
1080  LET L3=L+T1*L1
1090  PRINT "THE REGRESSION STATISTICS ARE AS FOLLOWS:"
1100  PRINT
1110  LET I5=ABS(I)
1120  IF I5-I=0 THEN 1150
1130  PRINT "THE EQUATION OF THE LINE IS:  Y=";L;"-";I5;"X"
1140  GO TO 1160
1150  PRINT "THE EQUATION OF THE LINE IS:  Y=";L;"+";I;"X"
1160  PRINT
1170  PRINT "WHERE THE SLOPE IS:  ";I
1180  PRINT "AND THE Y-INTERCEPT IS:  ";L
1190  PRINT "THE STANDARD ERROR OF THE REGRESSION IS:  (+ OR -) ";S
```

```
1200  PRINT "THE STANDARD ERROR OF THE SLOPE IS:  (+ OR -) ";I1
1210  PRINT "THE 95% C.L. FOR THE SLOPE ARE:  ";I2;"     ";I3
1220  PRINT "THE STANDARD ERROR OF THE INTERCEPT IS:  (+ OR -) ";L1
1230  PRINT "THE 95% C.L. FOR THE INTERCEPT ARE:  ";L2;"     ";L3
1240  PRINT "THE CORRELATION COEFFICIENT (R) IS:  ";N2
1250  PRINT "THE COEFFICIENT OF DETERMINATION (R**2) IS:  ";N1
1260  PRINT
1270  PRINT
1280  PRINT "DO YOU WANT MORE STATISTICS PRINTED?"
1290  PRINT "TYPE Y OR N."
1300  INPUT S5$
1310  IF S5$="N" THEN 1540
1320  PRINT
1330  PRINT
1340  PRINT "THE REGRESSION COMPUTATIONS HAVE PRODUCED THE FOLLOWING: "
1350  PRINT
1360  PRINT "THE MEANS OF X AND Y ARE:  ";J;"     ";K
1370  PRINT "THE SUM OF X IS:  ";A
1380  PRINT "THE SUM OF Y IS:  ";D
1390  PRINT "THE SUM OF X-SQUARED IS:  ";B
1400  PRINT "THE SUM OF Y-SQUARED IS:  ";E
1410  PRINT "THE SUM OF X*Y IS:  ";F
1420  PRINT "THE SUM OF SQUARES OF X IS:  ";G
1430  PRINT "THE SUM OF CROSS-PRODUCTS IS:  ";H
1440  PRINT "THE REGRESSION SUM OF SQUARES IS:  ";M
1450  PRINT "THE RESIDUAL SUM OF SQUARES IS:  ";O
1460  PRINT "THE TOTAL SUM OF SQUARES IS:  ";E1
1470  PRINT "THE REGRESSION MEAN SQUARE IS:  ";M
1480  PRINT "THE RESIDUAL MEAN SQUARE IS:  ";O
1490  PRINT
1500  PRINT "THE F-VALUE IS:  ";R
1510  PRINT "WITH 1 REGRESSION D OF F, AND"
1520  PRINT "A RESIDUAL D OF F OF :  ";P
1530  PRINT
1540  PRINT
1550  PRINT "DO YOU WANT 95% CONFIDENCE LIMITS?"
1560  PRINT "TYPE Y OR N."
1570  INPUT S6$
1580  IF S6$="N" THEN 1830
1590  PRINT
1600  PRINT "DO YOU WANT TO SPECIFY THE X VALUES?"
1610  PRINT "TYPE Y OR N."
1620  INPUT S7$
1630  IF S7$="Y" THEN 1660
1640  LET N4=N
1650  GO TO 1740
1660  PRINT
1670  PRINT "HOW MANY X VALUES DO YOU WANT TO ENTER?"
1680  INPUT N4
1690  PRINT "LIST THE EACH X VALUE BELOW."
1700  PRINT
1710  FOR C=1 TO N4
1720  INPUT X(C)
1730  NEXT C
1740  PRINT
```

```
 1750  PRINT "THE PREDICTED VALUES AND 95% C.L. OF Y ARE:"
 1760  PRINT
 1770  PRINT "GIVEN X VALUE   PREDICTED Y     LOWER Y        UPPER Y      ERROR"
 1780  FOR C=1 TO N4
 1790  REM T IS THE PREDICTED Y VALUE
 1800  LET T=L+I*X(C)
 1810  REM E5 IS THE STANDARD ERROR OF THE PREDICTED Y FOR THAT X VALUE
 1820  LET E5=SQR(Q*(1/N+((X(C)-J)¬2)/S))
 1830  REM Y1 AND Y2 ARE THE 95% C.L. AROUND THE PREDICTED Y
 1840  LET Y1(C)=T-(T1*E5)
 1850  LET Y2(C)=T+(T1*E5)
 1860  PRINT "     ";X(C);"   ";T;"      ";Y1(C);"   . ";Y2(C);"      ";E5
 1870  NEXT C
 1880  END
```

```
      REAL X,Y,B,R,A,N,MX,MY,VARX,VARY,COVXY,SLX2,SLY2,SLXY,RSS,ESS,TSSREG00010
      REAL RMS,EMS,F,R2,PERR2                                           REG00020
      INTEGER N,RDF,EDF,TDF,TX,TY                                       REG00030
      DIMENSION X(50),Y(50),YHAT(50),YRES(50)                           REG00040
      SUMX=0                                                            REG00050
      SUMY=0                                                            REG00060
      SUMXY=0                                                           REG00070
      SUMX2=0                                                           REG00080
      SUMY2=0                                                           REG00090
      I=1                                                               REG00100
      WRITE(5,12)                                                       REG00110
  12  FORMAT('THE DATA AS READ IN, BEFORE ANY TRANSFORMATION, ARE:')    REG00120
      WRITE(6,14)                                                       REG00130
  14  FORMAT(6X,'X',12X,'Y')                                            REG00140
      READ(2,*)TX,TY                                                    REG00150
   5  READ(2,*)X(I),Y(I)                                                REG00160
      IF(X(I).EQ.9999) GOTO 20                                          REG00170
      WRITE(6,16)X(I),Y(I)                                              REG00180
  16  FORMAT(F10.3,3X,F10.3)                                            REG00190
      IF(TX.EQ.1) X(I)=LOG(X(I))                                        REG00200
      IF(TX.EQ.2) X(I)=LOG(X(I)+1)                                      REG00210
      IF(TY.EQ.1) Y(I)=LOG(Y(I))                                        REG00220
      IF(TY.EQ.2) Y(I)=LOG(Y(I)+1)                                      REG00230
      SUMX=SUMX+X(I)                                                    REG00240
      SUMY=SUMY+Y(I)                                                    REG00250
      SUMXY=SUMXY+X(I)*Y(I)                                             REG00260
      SUMX2=SUMX2+X(I)*X(I)                                             REG00270
      SUMY2=SUMY2+Y(I)*Y(I)                                             REG00280
      I=I+1                                                             REG00290
      GO TO 5                                                           REG00300
  20  N=I-1                                                             REG00310
      WRITE(5,150)                                                      REG00320
      WRITE(6,22)                                                       REG00330
  22  FORMAT('THE DATA AFTER TRANSFORMATION, IF ANY, ARE:')             REG00340
      WRITE(6,14)                                                       REG00350
  24  FORMAT(6X,'X',12X,'Y')                                            REG00360
      DO 28 I=1,N                                                       REG00370
  28  WRITE(6,16)X(I),Y(I)                                              REG00380
  26  FORMAT(F10.3,3X,F10.3)                                            REG00390
      WRITE(5,150)                                                      REG00400
      MX=SUMX/N                                                         REG00410
      MY=SUMY/N                                                         REG00420
      SLX2=SUMX2-SUMX*SUMX/N                                            REG00430
      SLY2=SUMY2-SUMY*SUMY/N                                            REG00440
      SLXY=SUMXY-SUMX*SUMY/N                                            REG00450
      VARX=SLX2/(N-1)                                                   REG00460
      VARY=SLY2/(N-1)                                                   REG00470
      COVXY=SLXY/(N-1)                                                  REG00480
      B=SLXY/SLX2                                                       REG00490
      A=MY-B*MX                                                         REG00500
      RDF=1                                                             REG00510
      EDF=N-2                                                           REG00520
      TDF=N-1                                                           REG00530
      RSS=SLXY*SLXY/SLX2                                                REG00540
      ESS=SLY2-RSS                                                      REG00550
      TSS=SLY2
      RMS=RSS
      EMS=ESS/EDF
      F=RMS/EMS
      R2=RSS/TSS
      PERR2=100*R2
      WRITE(6,30)MX,MY
 30   FORMAT(' ',2X,'X MEAN=',F8.2,3X,'Y MEAN=',F8.2)
      WRITE(6,40)VARX,VARY,COVXY
 40   FORMAT('X VARIANCE=',F8.2,3X,'Y VARIANCE=',F8.2,3X,
     & 'XY COVARIANCE =',F8.2)
      WRITE(5,150)
      WRITE(6,150)
      WRITE(6,50)A,B
 50   FORMAT(' ','THE REGRESSION LINE IS Y=',F9.4,' + ',F9.4,'X')
      WRITE(5,150)
      WRITE(5,150)
      WRITE(6,60)
 60   FORMAT(' ','THE ANALYSIS OF VARIANCE TABLE IS:')
      WRITE(5,150)
      WRITE(5,70)
 70   FORMAT(' ','SOURCE',2X,'SUM OF SQUARES',2X,'MEAN SQUARE',2X,
     & 'F-STATISTIC')
      WRITE(6,80)RDF,RSS,RMS,F
 80   FORMAT(' ',I5,4X,F8.2,5X,F7.2,5X,F7.2)
      WRITE(5,90)EDF,ESS,EMS
 90   FORMAT(I5,4X,F8.2,5X,F7.2)
      WRITE(6,100)
 100  FORMAT(1X,'-----',4X,'---------')
      WRITE(5,110)TDF,TSS
 110  FORMAT(I5,4X,F8.2)
      WRITE(6,150)
      WRITE(5,120)R2,PERR2
 120  FORMAT(' ','R-SQUARED=',F6.5,3X,'PERCENT R-SQUARED=',F5.2)
      WRITE(6,150)
      WRITE(5,150)
      WRITE(6,180)
 180 FORMAT('Y-PREDICTEDS AND Y-RESIDUALS FOLLOW.')
      WRITE(5,150)
      WRITE(6,150)
      DO 130 I=1,N
      YHAT(I)=A+B*X(I)
 130  YRES(I)=YHAT(I)-Y(I)
      WRITE(5,150)
      WRITE(5,140)
 140  FORMAT('Y-PREDICTEDS',3X,'Y-RESIDUALS')
      DO 170 I=1,N
 170  WRITE(5,160)YHAT(I),YRES(I)
 160  FORMAT(F10.3,3X,F11.3)
 150  FORMAT(' ')
      STOP
      END
```

```
C                                                                       RSL00010
C   THIS PROGRAM CALCULATES A P VARIABLES-BY-P VARIABLES CORRELATION    RSL00020
C     MATRIX, ORDERS THE VARIABLES BY A DECREASING SUM OF               RSL00030
C     R SQUARED CRITERION, AND THEN IT CAN CONTINUE, SWEEPING           RSL00040
C     THE MATRIX OF ALL CORRELATIONS WITH THE FIRST VARIABLE, AND       RSL00050
C     REPEATING THE ORDERING AND SWEEPING PROCESS UNTIL ALL P           RSL00060
C     VARIABLES (OR SOME SPECIFIED SUBSET OF VARIABLES) HAVE BEEN       RSL00070
C     CHOSEN. THIS PROCEDURE AND THE ALGORITHM FOR DOING IT WAS         RSL00080
C     ORIGINALLY PROPOSED BY L. ORLOCI (1973; NATURE, LONDON 244:       RSL00090
C     371-373). PROGRAMS WRITTEN IN BASIC ARE GIVEN IN ORLOCI'S         RSL00100
C     1978 BOOK AND THE ORLOCI & KENKEL 1984 COURSE MANUAL (SEE THE     RSL00110
C     "BIBLIOGRAPHY" YOU WERE GIVEN FOR THE FULL REFERENCES). THIS      RSL00120
C     IMPLEMENTATION OF THE ALGORITHM IN FORTRAN IS BY R. H. GREEN.     RSL00130
C                                                                       RSL00140
C   THERE IS A CONTROL CARD, WHICH SHOULD BE FOLLOWED BY THE DATA       RSL00150
C     CARDS. THE N-BY-P DATA ARE ASSUMED TO BE IN "FREE FORMAT".        RSL00160
C                                                                       RSL00170
C   THE CONTROL CARD SHOULD HAVE ON IT (IN FREE FORMAT) THE             RSL00180
C           VARIABLES N, P, CYCLE, BRIEF, AND CPTC, WHERE:              RSL00190
C           (A) N = NUMBER OF SAMPLES (NOW DIMENSIONED FOR 200)         RSL00200
C           (B) P = NUMBER OF VARIABLES (NOW DIMENSIONED FOR 150)       RSL00210
C           (C) CYCLE = NUMBER OF VARIABLES TO BE CHOSEN                RSL00220
C           (D) BRIEF ( >0 IF PRINTOUT OF CORRELATION MATRICES IS       RSL00230
C              DESIRED, =0 OTHERWISE)                                   RSL00240
C           (E) CPTC = PERCENTAGE OF CORRELATION STRUCTURE TO BE        RSL00250
C              ACCOUNTED FOR                                            RSL00260
C                                                                       RSL00270
C                                                                       RSL00280
C                                                                       RSL00290
      DIMENSION X(600,150),RSQ(150,150),SUM(150),                       RSL00300
     &RSMJJ(150),R(150,150),SMR2(150),RANK(150),VARNO(150),FMTIN(20)    RSL00310
     &,FMTOUT(20),H(150),FRNK(150),DIAG(150),RI(150,130),TPCT(150)      RSL00320
      REAL X,RSQ,SUM,SUMSQ,SUMJK,RSMJK,RSMJJ,RSMJK2,R,SMR2,LARGE,RANK,  RSL00330
     &DIAG,RI,SIGN,TPCT,CPCT,FLIP                                       RSL00340
      INTEGER N,P,I,FMTIN,FMTOUT,J,K,L,VARNO,NUM,H,CYCLE,M,FRNK,BRIEF,  RSL00350
     &G,BEST,O                                                          RSL00360
      READ(2,*)N,P,CYCLE,BRIEF,CPCT                                     RSL00370
      FLIP=100-CPCT                                                     RSL00380
      WRITE(6,42)                                                       RSL00390
   42 FORMAT(' ')                                                       RSL00400
      WRITE(6,62)N,P,CYCLE,BRIEF,CPCT                                   RSL00410
   62 FORMAT(' N =',I4,6X,'P =',I4,5X,'CYCLE =',I3,6X,                  RSL00420
     &'BRIEF =',I2,6X,'CPCT =',F4.0)                                    RSL00430
      WRITE(6,*)' '                                                     RSL00440
      WRITE(6,20)                                                       RSL00450
   20 FORMAT('THE N SAMPLES-BY-P VARIABLES INPUT DATA ARE:')            RSL00460
      DO 30 I=1,N                                                       RSL00470
      READ(2,*)(X(I,J),J=1,P)                                           RSL00480
      WRITE(6,1030)(X(I,J),J=1,P)                                       RSL00490
 1030 FORMAT(8F10.3)                                                    RSL00500
   30 CONTINUE                                                          RSL00510
      WRITE(6,*)' '                                                     RSL00520
      DO 40 J=1,P                                                       RSL00530
      DO 50 I=1,N                                                       RSL00540
      SUM(J)=SUM(J)+X(I,J)                                              RSL00550
      SUMSQ=SUMSQ+(X(I,J))*(X(I,J))
   50 CONTINUE
      RSMJJ(J)=SUMSQ-((SUM(J))*(SUM(J))/N)
      SUMSQ=0.0
   40 CONTINUE
      DO 60 J=1,P
      DO 60 K=J,P
      DO 70 I=1,N
      SUMJK=SUMJK+(X(I,J))*(X(I,K))
   70 CONTINUE
      RSMJK=SUMJK-((SUM(J)*SUM(K))/N)
      SIGN=RSMJK/ABS(RSMJK)
      RSMJK2=RSMJK*RSMJK
      RSQ(J,K)=RSMJK2/(RSMJJ(J)*RSMJJ(K))
      R(J,K)=SQRT(RSQ(J,K))
      R(J,K)=SIGN*R(J,K)
      RSQ(K,J)=RSQ(J,K)
      R(K,J)=R(J,K)
      SUMJK=0.0
   60 CONTINUE
  170 IF(BRIEF.EQ.0)GO TO 22
      WRITE(5,*)' '
      WRITE(6,80)
   80 FORMAT('THE P-BY-P CORRELATION MATRIX IS:')
      DO 100 J=1,P
      WRITE(6,110)(R(J,K),K=1,P)
  110 FORMAT(15F8.4)
  100 CONTINUE
      WRITE(5,*)' '
      DO 111 J=1,P
      DO 111 K=1,P
  111 SMR2(J)=0.0
   22 DO 120 J=1,P
      DO 120 K=1,P
      SMR2(J)=SMR2(J)+RSQ(J,K)
      H(J)=J
  120 CONTINUE
      M=1
      J=1
  130 Q=P-J+1
      LARGE=SMR2(1)
      NUM=H(1)
      L=1
      IF(J.EQ.1)GO TO 1020
      DO 6 J=2,Q
      IF(SMR2(J).LE.LARGE)GO TO 6
      LARGE=SMR2(J)
      NUM=H(J)
      L=J
    6 CONTINUE
 1020 P=Q+J-1
      RANK(M)=LARGE
      VARNO(M)=NUM
      M=M+1
      H(L)=H(P-J+1)
      SMR2(L)=SMR2(P-J+1)
      J=J+1
      IF(J.LE.P) GO TO 130
      WRITE(6,135)
  135 FORMAT('THE VARIABLES (TOP) ARE ORDERED BY THE SUM-OF-RSQUARED
     &CRITERION (BOTTOM) :')
      WRITE(6,24)(VARNO(J),J=1,P)
      WRITE(6,23)(RANK(J),J=1,P)
   24 FORMAT(13I10)
   23 FORMAT(13F10.4)
      G=G+1
      FRNK(G)=VARNO(1)
      BEST=VARNO(1)
      DO 155 J=1,P
      DO 155 K=1,P
      IF(J.NE.K)GO TO 155
      DIAG(G)=DIAG(G)+R(K,J)
  155 R1(K,J)=R(K,BEST)*R(BEST,J)/R(BEST,BEST)
      DO 160 J=1,P
      SMR2(J)=0.0
      DO 160 K=J,P
      R(J,K)=R(J,K)-R1(J,K)
      RSQ(J,K)=R(J,K)*R(J,K)
      R(K,J)=R(J,K)
      RSQ(K,J)=RSQ(J,K)
  160 CONTINUE
      WRITE(6,*)' '
      WRITE(6,165)G
  165 FORMAT(I4,1X,'VARIABLES & THEIR CORRELATIONS HAVE',
     &' BEEN REMOVED.')
      WRITE(6,*)' '
      TPCT(G)=100*(DIAG(G)/DIAG(1))
      IF(G.EQ.P)GO TO 900
      IF(G.EQ.CYCLE)GO TO 950
      IF(TPCT(G).LT.FLIP)GO TO 965
      GO TO 170
  900 WRITE(6,*)' '
      WRITE(6,910)
  910 FORMAT('THE CORRELATION MATRIX IS NOW EXHAUSTED.')
      GO TO 960
  965 WRITE(6,*)' '
      WRITE(6,970)CPCT
  970 FORMAT(F4.0,1X,'% OF THE CORRELATION STRUCTURE HAS BEEN',
     &' ACCOUNTED FOR.')
      GO TO 960
  950 WRITE(6,955)CYCLE
  955 FORMAT('-',I4,1X,'CYCLES HAVE BEEN DONE.')
  960 WRITE(6,*)' '
      WRITE(6,920)
  920 FORMAT('THE BEST ORDER IN WHICH TO CHOOSE VARIABLES, FOR',
     &' MAXIMUM INFORMATION ABOUT STRUCTURE IN THE DATA SET, IS:')
      WRITE(6,930)(FRNK(I),I=1,G)
  930 FORMAT(16I8)
      WRITE(6,*)' '
      WRITE(6,935)
  935 FORMAT('THE TRACES ASSOCIATED WITH THE RESIDUAL',                 RSL01660
     &' CORRELATION MATRICES ARE (BEGINNING WITH 0 VARIABLES',          RSL01670
     &' REMOVED):')                                                     RSL01680
      WRITE(6,940)(DIAG(I),I=1,Q)                                       RSL01690
  940 FORMAT(16F8.3)                                                    RSL01700
      WRITE(6,*)' '                                                     RSL01710
      WRITE(6,975)                                                      RSL01720
  975 FORMAT('THE TRACES, AS A PERCENTAGE OF P, ARE:')                  RSL01730
      WRITE(6,980)(TPCT(I),I=1,Q)                                       RSL01740
  980 FORMAT(16F8.1)                                                    RSL01750
      STOP                                                              RSL01760
      END                                                               RSL01770
```

```
FILE: PLOT     FORTRAN  A1  VM/SP - CONVERSATIONAL MONITOR SYSTEM                    PAGE 001

C  PROGRAM PLOT
C
      DIMENSION V(50,50)
      CHARACTER*1  YES,XNO,PLS(14),PLOTT(65)
      DATA YES, XNO /'Y','N'/
      DATA PLS /'0','1','2','3','4','5','6','7','8','9',
     X  ' ','-','I','.'/
      COMMON  MOR(50,60),PLOTT
      J=1
      WRITE(6,*)'THE DATA AS READ IN ARE:'
   10 READ(2,*)(V(I,J),I=1,2)
      IF(V(1,J).EQ.9999) GOTO 15
      WRITE(6,*)V(1,J),V(2,J)
      J=J+1
      GO TO 10
   15 M=J-1
      WRITE(6,*)' '
      IWI=60
      I=1
      J=2
      DO 1  IGH=1,50
      DO 1  JGH=1,IWI
    1 MOR(IGH,JGH)=10
      WI=IWI
      ICODEX=0
      ICODEY=0
      XMIN=V(1,1)
      XMAX=XMIN
      DO 20  K=2,M
      IF(V(1,K).GT.XMAX) XMAX=V(I,K)
      IF(V(1,K).LT.XMIN)  XMIN=V(I,K)
   20 CONTINUE
      YMIN=V(2,1)
      YMAX=YMIN
      DO 21  K=2,M
      IF(V(2,K).GT.YMAX)  YMAX=V(2,K)
      IF(V(2,K).LT.YMIN)  YMIN=V(2,K)
   21 CONTINUE
      XVAL=XMAX-XMIN
      YVAL=YMAX-YMIN
      UNX=XVAL/WI
      UNY=YVAL/20.0
C  INSERT AXES IF POSSIBLE
      IF(YMIN.GT.0.0.OR.YMAX.LT.0.0)  GO TO 50
      IX=19.0-(-YMIN+UNY)/UNY
      IF(IX.GT.60)  IX=50
      IF(IX.LT.1)  IX=1
      ICODEX=1
      DO 35 L=1,IWI
   35 MOR(IX,L)=11
   50 IF(XMIN.GT.0.0.OR.XMAX.LT.0.0)  GO TO 59
      IY=(-XMIN+UNX)/UNX
      IF(IY.GT.IWI) IY=IWI
      IF(IY.LT.1)  IY=1
      ICODEY=1
      DO 52 L=1,20
   52 MOR(L,IY)=12
   53 TICX=0.1
      TICY=0.1
      IF(XVAL.GT.2)  TICX=0.5
      IF(XVAL.GT.12)  TICX=1
      IF(XVAL.GT.40)  TICX=10
      IF(YVAL.GT.2)  TICY=0.5
      IF(YVAL.GT.12)  TICY=1
      IF(YVAL.GT.40)  TICY=10
      START=0.0
      IF(ICODEX.NE.1)  GO TO 64
      IF(ICODEY.NE.1.AND.XMAX.LT.0.0)  GO TO 62
   61 START=START+TICX
      IF(START.GT.XMAX) GO TO 62
      IF(START.GT.XMAX)  GO TO 62
      LX=(START-XMIN+UNX)/UNX
      IF(LX.GT.IWI) LX=IWI
      IF(LX.LT.1)  LX=1
      MOR(IX,LX)=12
      GO TO 61
   62 START=0.0
      IF(ICODEY.NE.1.AND.XMIN.GT.0.0)  GO TO 64
   63 START=START-TICX
      IF(START.LT.XMIN)  GO TO 64
      IF(START.GT.XMAX)  GO TO 63
      LX=(START-XMIN+UNX)/UNX
      IF(LX.LT.1)  LX=1
      IF(LX.GT.IWI)  LX=IWI
      MOR(IX,LX)=12
      GO TO 63
   64 START=0.0
      IF(ICODEY.NE.1) GO TO 74
      IF(ICODEX.NE.1.AND.YMAX.LT.0.0)  GO TO 70
   65 START=START+TICY
      IF(START.GT.YMAX)  GO TO 70
      IF(START.LT.YMIN)  GO TO 65
      LY=18.0-(START-YMIN+UNY)/UNY
      IF(LY.LT.1)  LY=1
      IF(LY.GT.18)   LY=18
      MOR(LY,IY)=11
      GO TO 65
   70 START=0.0
      IF(ICODEX.NE.1.AND.YMIN.GT.0.0)  GO TO 74
   71 START=START-TICY
      IF(START.LT.YMIN)  GO TO 74
      IF(START.GT.YMAX)  GO TO 71
      LY=18.0-(START-YMIN+UNY)/UNY
      IF(LY.GT.18)  LY=18
      IF(LY.LT.1)  LY=1
      MOR(LY,IY)=11
      GO TO 71
   74 IF(ICODEY.NE.1) TICY=0.0
      IF(ICODEX.NE.1)  TICX=0.0
      WRITE(6,110) I,J,XMIN,XMAX,UNX,TICX,YMIN,YMAX,UNY,TICY
```

XX

```
 110  FORMAT (//3X,'HORIZONTAL AXIS IS DIMENSION',I3/
     X 3X,'VERTICAL AXIS IS DIMENSION',I5/3X///10X,'HORIZONTAL AXIS'/
     X/3X,'MINIMUM VALUE=',F15.5/3X,'MAXIMUM VALUE=',F15.5/3X,
     X    'SCALING UNIT =',F15.5/3X,'ONE TICK=',F10.3///10X,
     X'VERTICAL AXIS'//3X,'MINIMUM VALUE=',F15.5/3X,'MAXIMUM VALUE=',
     XF15.5/3X,'SCALING UNIT =',F15.5/3X,'ONE TICK=',F10.3//3X,
     X   'OVERLAPPING OBJECTS (NOT PLOTTED)'//3X,'ID.NUMBER',3X,
     X    'COORDINATES'/)
      DO 100 L=1,M
      X=V(I,L)
      Y=V(J,L)
      IX=(X-XMIN+UNX)/UNX
      IY=18.0-(Y-YMIN+UNY)/UNY
      IF(IX.GT.IWI)  IX=IWI
      IF(IX.LT.1)  IX=1
      IF(IY.GT.18)  IY=18
      IF(IY.LT.1)  IY=1
      IF(L.GE.100)  GO TO 790
      IF(L.GE.10)  GO TO 780
      IF(MOR(IY,IX).LE.9)  GO TO 800
      MOR(IY,IX)=L
      GO TO 100
 780  IF(IX.EQ.IWI)  IX=IX-1
      IF(MOR(IY,IX).LE.9.OR.MOR(IY,IX+1).LE.9)  GO TO 800
      MOR(IY,IX)=L/10
      MOR(IY,IX+1)=L-(L/10)*10
      GO TO 100
 790  IF(IX.EQ.IWI)  IX=IX-2
      IF(IX.EQ.IWI-1)  IX=IX-1
      IF(MOR(IY,IX).LE.9.OR.MOR(IY,IX+1).LE.9.OR.MOR(IY,IX+2).LE.9)
     X   GO TO 800
      MOR(IY,IX)=L/100
      MOR(IY,IX+1)= (L-(L/100)*100)/10
      MOR(IY,IX+2)=L-(L/10)*10
      GO TO 100
 800  WRITE(6,801) L,X,Y
 801  FORMAT (I9,5F15.5)
 100  CONTINUE
 111  FORMAT (3X,125A1)
 115  FORMAT (1H1)
      WRITE(6,115)
      DO 113 JJ=1,IWI
 113  PLOTT(JJ)=PLS(14)
      WRITE(6,111) PLS(14), (PLOTT(JJ),JJ=1,IWI), PLS(14)
      DO 200 K=1,18
      DO 150 L=1,IWI
 150  PLOTT(L)=PLS(MOR(K,L)+1)
 200  WRITE(6,111)PLS(14) ,(PLOTT(KL), KL=1,IWI), PLS(14)
      DO 201  JJ=1,IWI
 201  PLOTT(JJ)=PLS(14)
      WRITE(6,111) PLS(14), (PLOTT(JJ),JJ=1,IWI), PLS(14)
      STOP
      END
```

APPENDIX III - COUNTRY REPORTS

III.1 Indonesia

THE APPLICATION OF COMPUTER AT THE INDONESIAN INSTITUTE OF SCIENCES AND THE UNIVERSITIES IN INDONESIA

Tri Surja Kreshnawati
(LIPI Jakarta)

S. Djalal Tandjung
(UGM Yogyakarta)

Introduction

LIPI is a Government body which provides guidance in the field of scientific and technological research. It reports directly to the President of the Republic of Indonesia.

LIPI has ten national research institutions situated in Jakarta, Bogor, Bandung, and Serpong which are conducting research in the natural, technological and social science. There is also a National Scientific Documentation Centre.

The national research institution administered by LIPI are:

- National Biological Institute
- National Institute of Oceanology
- National Institute of Geology and Mining
- National Institute for Chemistry
- National Institute for Physics
- National Institute for Metalurgy
- National Institute for Electrotechniques
- National Institute for Instrumentation
- National Institute for Economic & Social Research
- National Institute for Cultural Studies

Computer in LIPI

The rapid advance of Science and Technology in the last two decades can be attributed mostly to the intelligent use of computers in data handling and analysis.

A computer is used because it does certain task and ability better and more efficiently than mankind. The characteristics of this machine are speed and capacity to handle large volumes of data in a very short time. It is far from exaggeration that computers in the advancement of Science and Technology are indispensable. Each of the national research institute use computers for R & D activities. In this case, we describe one of the institute is National Biological Institute, and in addition some information on the usage of computers in higher education Institutions in Indonesia.

National Biology Institute - LIPI

The National Biology Institute has an Apple II computer with 48 K. capacity and a silent type printer. The printer can print 132 characters and has the capacity to print graphics.

Available computer programmes are as follows:

1. Visifile for information on management data.
2. Visitrend for analysis and graphics.
3. Visicalc for genetic pool collection.
4. Abstat for statistic analysis.
5. Utilities for visifile.
6. DOS 3.3.

At the present time the National Biology Institute has computerised diskettes for:

1. Documental ethnobotany collection.
2. Botanical Garden collection.
3. Genetic pool garden collection.
4. Herbarium Bogoriense collection.
5. Zoology Museum collection.
6. Ecology research.
7. Taxonomy.

The data discrete programme storage specifications are:

a. One data sheet.
b. 24 column for one file, with 232 characters.

Example :

Ethnobotany collection

Registration number collector collector number date
location region Name of thing material plant useful

The steps are:

1. formulate the format.
2. data entry.
3. data storage.

The data storage can be used at any time. Based on the example, the data can be processed as it is needed, for example:

- What kind of matter at the vitrin 7
- What kind of collection from West Kalimantan
- For what purpose the rottan are used, etc.

Botanical Garden Collection

Family Species type/variety Island Location Altitude
No. of plan Date of plan Herbarium Blooming Fertilization

From the data entry can be used for:

- What is the number of Herbarium material.
- What kind of collection from Sumatra.
- How many *Pterospermum javanicum* is grown.
- When was the *Eucalyptus alba* planted.

Plant Ecology

Plotting species family diameter basal area unbranched trunk
Total high topography soil

The data can be used to determine:

- What kind of species has diameter of 50cm.
- What kind of species belong to the group of Myrtaceae.

Herbarium specimen

Registration Number Family Number Species Local name Island location height habitat collector Number Date

- How many genus and species are kept in the Herbarium Bogoriense.
- What species are collected from Sumatra Barat.
- What species are found only at high elevation e.g. 750 meters above sea level.
- What are the Orchidacea family fund in Sumatra.

The computer is also used for finalised legume data sheet as below.

Legume data sheet

Collection Data

Accession number Scientific name Local/English name Collection number Collection date Collection site Material collected Occurrence Uses

Evaluation Data

Habitat Plant type Life duration leaf type leaflet shape Flower colour Pod type Pod shape Pod texture Pod colour Seed shape Seed colour Tuber Flowering time Age of first flower Pod setting Pod length Number of seeds per pod 100 seed weight Disease resistance Pest tolerance

Additional notes

Those are examples of several usages of computer in the National Biology Institute of LIPI.

The Application of Computer in the Universities in Indonesia

This is not an official information based on any research or survey. To the authors knowledge, some big universities such as Gadjah Mada University in Yogyakarta and Indonesia University in Jakarta have used computers in their work.

Gadjah Mada University has a computer center, which can be used for education and research by students and teaching staffs. Student from Faculty of Mathematics and Science have to take subjects on computer. Other students from other faculty use the computer as it is needed for data processing of their research. So far, computers have been used in many universities for education and research.

While the new generation of students (started with the year 1975) have the ability to operate the computers, their professors are left far behind, because in their age, when they were students, they did not get any computer training. Now the professors have to catch up today's computer technology.

Conclusion

LIPI and higher education institutions in Indonesia have started using computers in their work. More staff have to be trained to handle and be familiar with the computer.

III.2 Malaysia

THE STATUS OF COMPUTER HARDWARE AND SOFTWARE FACILITIES AND THE USE OF COMPUTERS FOR RESEARCH IN ENVIRONMENTAL BIOLOGY IN MALAYSIA

Kam Suan Pheng
(Universiti Sains Malaysia)

Roslan bin Ismail
(Forest Research Institute)

A. COMPUTER HARDWARE FACILITIES

A number of universities and research institutions in Malaysia are involved in biology and applied biology, and most of these institutions are equipped with some model of mainframes. The following table summarises the mainframes available at the various institutions, to the best of our knowledge. Therefore, this list is not exhaustive.

Institution	Mainframes and superminis
FRI	Data General Eclipse S140
MARDI	IBM
PORIM	HP 3000
RRIM	HP 3000
UM	IBM
UKM	IBM, PRIME
UPM	UNIVAC
USM	IBM 4331 & IBM 4381
UTM	IBM
IMR	IBM

FRI	Forest Research Institute
MARDI	Malaysian Agricultural Research and Development Institute
PORIM	Palm Oil Research Institute of Malaysia
PRIM	Rubber Research Institute of Malaysia

UM	Universiti Malaya (University of Malaya)
UKM	Universiti Kebangsaan Malaysia (National University of Malaysia)
UPM	Universiti Pertanian Malaysia (Agricultural University of Malaysia)
USM	Universiti Sains Malaysia (University of Science, Malaysia)
UTM	Universiti Teknologi Malaysia (University of Technology Malaysia)
IMR	Institute for Medical Research

Apart from the mainframes and super-minis, some of the institutions have mini-computers and micro-computers. Besides the government and quasi-government bodies listed above, there are also private research laboratories, such as those associated with the plantation and pesticide companies, which utilise computers in their research activities.

B. SOFTWARE AND PROGRAMS

The main frames available in the institutions listed above are normally used for large projects with big data sets. Programs are either written, in FORTRAN or BASIC, or software packages such as SAS, SPSS, BMDP, and GENSTAT are used. In a number of research organisations we know, the SAS package is preferred by scientists, especially biologists, because of its greater utility for analysing scientific data.

Specifically, we know of the use of computers for research purposes for the two institutions which we come from:

1. <u>Forest Research Institute</u>

a. Prefelling and post-felling inventorisation

b. Growth and yield studies

c. Numerous other aspects of forestry research

III.3 Philippines

STATUS OF COMPUTER APPLICATIONS TO ECOLOGICAL RESEARCHES/PROJECTS WITHIN THE MAB PROGRAMME IN THE PHILIPPINES

Christian P. Dizon
Jesus P. Bayrante
(Man & the Biosphere Inter-Agency Committee on Ecological Studies)

The Man and the Biosphere Inter-Agency Committee on Ecological Studies (MAB-ICES) in the Philippines implements its programme of research through cooperation and collaboration with its fourteen (14) government member-agencies and three (3) cooperating agencies which are all involved in resource management (see Attachment). As of March 1985, the MAB-ICES has continued to undertake at least nine (9) national ecological field projects for cooperative research all within the framework of the MAB Programme. Such researches are classified under UNESCO-MAB's international themes (i.e. forest areas, coastal zones, pollution, energy utilization, environmental impact assessment and biosphere reserves).

Quantitative analyses of significant ecological/environmental data generated from the various researches/projects within the MAB programme using the computers have not been widely used due to the lack of computer hardwares/machine/gadgets and technical personnel with the proper training who could easily process the data with the computers using the various quantitative/statistical packages being utilized by other countries.

An inventory of the kinds/types of mainframe computer hardwares and the corresponding softwares used by MAB member-agencies in which MAB researchers/scientists could have access to, revealed that there are more or less five (5) agencies with the mainframe computer hardwares. These hardwares are of the IBM (e.g. IBM 1130, etc.) and in most cases, the FORTRAN language is used.

In terms of micro or minicomputers, the Apple II-E, Apple II-Plus and IBM PC Compatible are the most common. The operating systems utilized are the TRS-DOS by Tandy, CP/M, COMMODORE DOS and MS-DOS.

Based on the foregoing, there is a need to expose the reseachers/scientists within the MAB Programme in the Philippines to the current and perhaps advanced statistical packages using the above mentioned computers especially the micros. With this, data handling, storage, processing and analysis would be facilitated.

MAB PHILIPPINES GOVERNMENT MEMBER AGENCIES

1. Bureau of Plant Industry
2. Bureau of Animal Industry
3. Bureau of Soils
4. Bureau of Lands
5. Bureau of Mines and Geosciences
6. Bureau of Fisheries and Aquatic Resources
7. Bureau of Forest Development
8. Bureau of Coast and Geodetic survey
9. National Institute of Science and Technology
10. National Museum
11. Philippine Atmospheric, Geophysical, and Astronomical Services Administration
12. National Irrigation Administration
13. Ministry of Public Works & Highways
14. Philippine Coast Guard

COOPERATING AGENCIES

1. National Pollution Control Commission
2. Forest Research Institute
3. National Water Resources Council

III.4 Singapore

THE USE OF COMPUTERS IN THE SCHOOLS OF SINGAPORE AND IN THE DEPARTMENT OF ZOOLOGY, NUS

Choo Bee Li
(National University of Singapore)

Tan Siok Cheng
(Curriculum Development Institute of Singapore)

The Singapore government started promoting computer awareness in the schools in 1980. Ample funds were allocated to be various educational institutions to purchase computers and train personnel to meet the demands of a sophiscated technological era. This report touches on the present computer situation in the various educational institutions of Singapore.

TEACHER TRAINING

The Institute of Education has a computer laboratory and conducts computer literacy courses for primary school teachers. It also has two terminals attached to the mainframe computer at the Ministry of Education for the use of its staff, trainee teachers and M Ed students.

The Curriculum Development Institute of Singapore's computer department has a computer laboratory which is well stocked with many IBM and a few Apple micro-computers. It conducts computer literacy lessons in BASIC to secondary school teachers and courses on the use of various software packages such as Logo for primary school teachers and Superpilot and dBaseII for secondary school teachers.

THE SCHOOLS

The junior colleges have their own computer laboratories and student can opt to take computer science as an 'A' Level examination subject.

Each secondary school in Singapore has three to ten micro-computers. Some SAP (Special Assistance Plan) schools have as many as twenty-five. These computers belong to the schools' computer clubs which normally conduct computer appreciation courses for their members. Some SAP schools give compulsory computer literacy lessons to their students.

The staff of some schools use computers to compute their school records and examination results.

Most of the primary schools do not have computers. The CDIS CAI (Computer Assisted Instruction) Project Team is preparing, for a start, a computer laboratory in one primary school. It should be ready by this July. It will have a mainframe computer and twenty-four on-line terminals. The project team intends to introduce CAI packages in mathematics, mainly of the drill and practice type to the weaker students in the primary schools.

THE DEPARTMENT OF ZOOLOGY, NUS

The Department of Zoology of the National University of Singapore has about eighteen micro-computers and two mainframe terminals. The micro-computers are used mainly for teaching. For example, the fisheries courses for third and honours year students make extensive use of the computers. As micro-computers have small memory spaces, they are only used to analyse simple and small data sets. The micro-computer is also used to catalog the specimens in the Zoological Record Collection of the department.

The mainframe terminals with their more powerful software packages such as SAS and Minitab are well utilised by the staff and students of the department. The software packages can perform complex data manipulations such as multivariate statistical analysis.

III.5 Thailand

A REPORT FROM THE PARTICIPANTS OF THAILAND

Santad Koompalum
(National Environment Board)

Rojchai Satrawaha
(Khon Kaen University)

Air and noise pollution section, environmental quality standard division, office of the National Environment board (ONEB) is responsible for technical data, policy determination and management of air and noise pollution in Thailand.

In a field of technical data, involved the monitoring of ambient air. There are 8 monitoring stations located in Bangkok area and a mobile monitoring unit is used to monitor air quality in other main cities and other areas which have air pollution problems. Other data include air pollution emission from motor vehicles. From industrial plants, noise and vibration data. Most of the analyzed data are assessed to provide input to the special committee for the determination of air quality standards for ambient air quality. Emission from motor vehicles and emission from industries. Some of the data is also used as information for other government unit and public sector which are concerned with air and noise pollution problems and control.

Raw data are collected continuously by automatic air pollutant analyzers for carbon monoxide, hydrocarbons, sulfur dioxide, oxides of nitrogen, oxidants and suspended particulate matter as charts from recorders and as data cassette tape recorder from dataloggers. Other raw data are from the meterological department. Traffic volume and industrial information are also obtained.

There are three microcomputer systems used in ONEB at the present. Now a Fujitsu micro-8 computer system and data cassette recorder are used in air and noise pollution section. Some of the software are developed in F-Basic language. Other packages include DBase II, Supercalc, Wordstar and Fortran-86 (16 Bit) under CP/M. There are two programmers with B.Sc. in statistics who operate the computer.

The Apple IIe and victor 4 system are used in water quality section and solid waste section respectively, ONEB is about to purchase the two 16 bit microcomputer systems under eastern seaboard project and ONEB also plans to have a mini computer to be used as environmental information center and data base for Thailand in the near future.

At the Faculty of Science, Khon Kaen University, there are fifteen Apple II microcomputers which can be used by the university staff. Environmental biologists usually do not have much background on computers. Analysis of biological data is mostly done with assistance from the mathematics and statistics department staff. However, Khon Kaen University has a plan to set up a computer center with mainframe facilities in the near future.

APPENDIX IV - LIST OF PARTICIPANTS

Jesus P. BAYRANTE
Man and the Biosphere Inter-Agency
Committee on Ecological Studies (MAB-ICES)
4th Floor, Asia Trust Bank Building
1424 Quezon Avenue
Quezon City
Philipines

Tran Thanh BINH
Institute of Physics
Ngihia Do Liem
Hanoi
Vietnam

Bee Li CHOO
Department of Zoology
National University of Singapore
Kent Ridge
Singapore 0511.

Christian P. DIZON
Man & the Biosphere Inter-agency Committee on
Ecological Studies (MAB-ICES)
4th Floor Asia Trust Bank Bldg
1424 Quezon Avenue
Quezon City
Philippines

Roslan bin ISMAIL
Forest Research Institute
Kepongi
Selangor
Malaysia

Suan Pheng KAM
School of Biological Sciences
Universiti Sains Malaysia
Penang
Malaysia

Jeong Gyu KIM
Korea National Environmental Protection Institute (NEPI)
#280-17 Bulkwang-dong, Eunpyung-ku, 122
Seoul
Korea

Santad KOOMPALUM
(National Environment Board)
60/1 Soi Pibol Wattana Bld
Rama VI Rd
Samsen
Bangkok
Thailand

Tri Suria Kreshnawati MOEIS
Indonesia Institute of Sciences
Bureau of Coordination & Science Policy
Widya Graha LIPI
jl Gatot Subroto
Jakarta
Indonesia

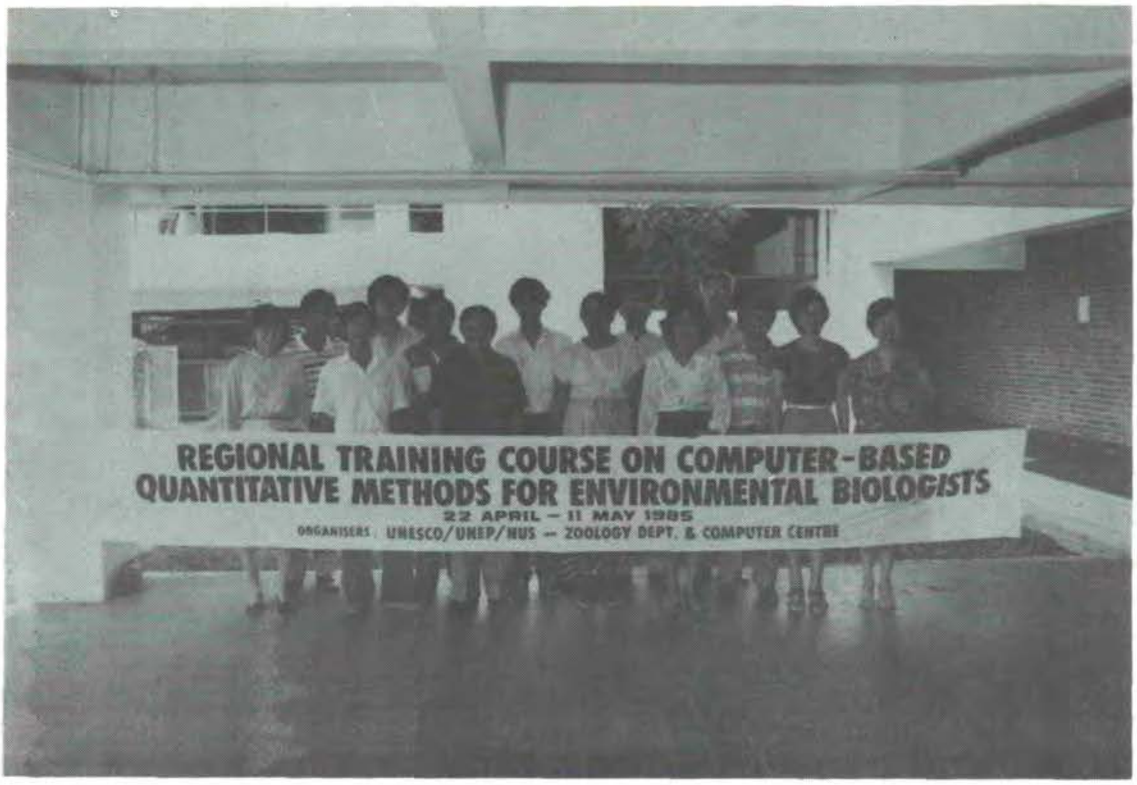
Sinapi MOLI
Department of Education
Malifa
Apia
Western Samoa.

Rojchai SATRAWAHA
Department of Biology
Faculty of Science
Khon Kuen University
Khon Kaen
Thailand 4002

Siok Cheng TAN
Curriculum Development Institute of Singapore
465-E Bukit Timah Rd
Singapore 1025.

S. Djalal TANDJUNG
Gadjah Mada University
Bulaksumur
Yogyakarta
Indonesia

Bennan WANG
The Commision for Integrated Survey of Natural Resources
Academia Sinica
#917 Building Datun Road
Beijing
P.R. China


REGIONAL TRAINING COURSE ON COMPUTER-BASED
QUANTITATIVE METHODS FOR ENVIRONMENTAL BIOLOGISTS
22 APRIL – 11 MAY 1985
ORGANISERS : UNESCO/UNEP/NUS — ZOOLOGY DEPT. & COMPUTER CENTRE

From left to right: Miss Bee Li CHOO, Mr Christian P. DIZON, Mr Santad KOOMPALUM, Mr Jeong Gyu KIM, Dr Roslan bin ISMAIL, Dr S. Djalal TANDJUNG, Mr Tran Thanh BINH, Mrs Sinapi MOLI, Dr Rojchai SATRAWAHA, Mrs Tri Suria Kreshnawati MOEIS, Prof Roger H. Green, Mr Jesus P. BAYRANTE, Dr Suan Pheng KAM, Miss Siok Cheng TAN, Mr Bennan WANG (absent).