MCH/91.14
Distr.: LIMITED

PHYSICAL GROWTH AND PSYCHOSOCIAL DEVELOPMENT OF CHILDREN: MONITORING AND INTERVENTIONS

GROSTAT

A PROGRAMME FOR ESTIMATING AGE RELATED CENTILES USING PIECEWISE POLYNOMIALS

Jon Rasbash
Huiqi Pan
Harvey Goldstein

Institute of Education, University of London

CHILD HEALTH AND DEVELOPMENT
MATERNAL AND CHILD HEALTH AND FAMILY PLANNING
WORLD HEALTH ORGANIZATION
GENEVA, 1992

WORLD HEALTH ORGANIZATION

ORGANISATION MONDIALE DE LA SANTE

DISTR. : LIMITED
DISTR. : LIMITEE

MCH/91.14
ENGLISH ONLY

PHYSICAL GROWTH AND PSYCHOSOCIAL DEVELOPMENT OF CHILDREN
MONITORING AND INTERVENTIONS

GROSTAT
A PROGRAMME FOR ESTIMATING AGE RELATED CENTILES USING PIECEWISE POLYNOMIALS

John Rasbash
Huiqi Pan
Harvey Goldstein

Institute of Education, University of London

CHILD HEALTH AND DEVELOPMENT
MATERNAL AND CHILD HEALTH AND FAMILY PLANNING
WORLD HEALTH ORGANIZATION
GENEVA, 1991

Table of Contents

PREFACE

We are very grateful to Bob Prosser for providing extracts from the ML3 software manual (Prosser et al, 1990). We are also most grateful to Richard Lansdown for help with preparing this manual.

1 INTRODUCTION

This document describes the second release of GROSTAT, a package for estimating age related centiles (Rasbash, 1987). The theory underpinning GROSTAT is outlined in Healy et al (1988), Goldstein & Pan (1991) and Pan & Goldstein (1990) this is implemented in GROSTAT. The kernel of Healy's NANOSTAT software (Healy, 1987), a general-purpose statistical package, provides the foundation and many of the utility features of the program. The first version of GROSTAT was written by Jon Rasbash.

The present version of the software is based upon the theory described in Goldstein and Pan (1991), and specifically includes the following features. A long age range may now be fitted satisfactorily for centiles curves by use of smoothly joining polynomials: Several populations or subgroups may be fitted simultaneously: the percentiles may be differentially weighted during estimation: individual data points may be differentially weighted and graphical goodness of fit tests may be used.

In this document, Section 1 describes the installation of GROSTAT and is followed by sections showing how to input data and carry out the necessary analyses. A HELP facility for the commands is available in the program.

2 An apology.

Our development priority with GROSTAT II has been to implement the new methodology presented in Goldstein & Pan(1989). Consequently we have been using a rather skeletal command interface. The procedures are relatively complex and therefore require a powerful and amenable command interface - we are well aware of the package's shortcomings in this area. We apologize in advance for any frustration this may cause. It is our intention to design a more accessible interface if the basic package proves sufficiently popular.

3 Installation

GROSTAT II uses a low level graphics package called GSX-86. Either Hercules or IBM CGA/EGA/VGA graphics cards are needed to use the program's high-resolution graphics capabilities. Two forms of the program are available: one requires 512k of RAM and the other 640k of RAM. Both need a hard disk.

You are required to create a subdirectory-call it c:\grostat, say-and move to it using the CD command. Place the program diskette in your disk drive A and type

copy a:*.*

Users of IBM CGA/EGA graphics card, type

install ibm

Users of Hercules graphics card should first copy the file MGC.EXE to your root directory and use an editing program to put the commands

mgc full

into your autoexec.bat file, then type

install herc

If you do not know what kind of graphics card you have, you could try both and see which works. If you experience any difficulties with the installation of the program, please contact us.

To run the program, in the subdirectory c:\grostat, type

gs

The -> prompt will appear on the screen.

To operate on data, type commands at this prompt. To exit from the program at the prompt, type

STOP

4 Inputing Data

This section simply describes the fundamentals about the software-the worksheet and the structure of commands; reading in data from a file and input vectors; checking the data.

4.1 The Worksheet

Data are stored by the program in a memory segment referred to as *worksheet* which has 100 columns and 100 boxes numbered from 1 to 100 respectively. Values for a given variable are contained in a single column, for example, C1. A single number can be held in a box, such as in K50. You can refer to consecutive columns as, for example, C1-C10 and boxes as K30-K35. For a microcomputer of 512k memory the worksheet is 25000 spaces and for 640k is 50000 spaces.

4.2 Commands and Notation

Every command of GROSTAT II must begin with a key word. This is followed by identifiers of the columns or boxes or by numbers. You may include explanatory words if you like. For example,

SUBT C1 C2 C3

is equivalent to

SUBTract C1 scores from scores in C2 and put the differences into C3

and

MULT N C1 C2

is equivalent to

MULTiply C1 by the value of N and put the product into C2

The symbols M and N in command expressions means either an ordinary number(such as 2 or 2.5) or a box indicator. The P in command espressions means a proportion (such as 0.1). The symbol K in a command expresses a box indicator and C a column.

The HELP facility is available to assist you in using the correct syntax for each command. Typing

HELP XXXX

(where XXXX is the name of a command) shows a brief note about the command's format.

4.3 Reading in Data from a File

Two modes of data-unformatted and formatted-can be read in by the command DREA with different syntaxes.

Suppose the WEIGHT.DAT is an unformatted file containing three columns of number. An unformatted data file is a file where the data columns are separated by one or more spaces. To read all columns from the file , type

DREAd C1-C3

You will be asked to type in the name of the file, and your response may refer to a drive indicator and /or path name as in C:\GROSTAT\DATA\WEIGHT.DAT.

For a formatted data, it is convenient to read in data using a FORTRAN-style format statement. This makes it possible to read data without separators between the items, to skip over some of the items in a record, and to read data records consisting of more than one line. The syntax is

DREAd N C...C

where N is either 1 or 2, the number of lines of format you will be asked to type.

Suppose the HEIGHT.DAT is a formatted file in the current directory including the items of id, age, birthday, height, etc. To read in id, age, height with one line of format, the work session is as follows

DREAd 1 C1-C3
Type file name
HEIGHT.DAT
Type 1 line(s) of format
(F3.0, 1x, F5.2, 8x, F6.2)
NAME C1 'ID' C2 'AGE' C3 'HEIGHT'

4.4 Inputing Vectors

Suppose that you will input a vector(1 3 3 5) into C1, you can type

SET C1
1 3 3 5

Note that you should type the numbers of the vector on the next line with spaces between them, going on to further lines as required and do not split a number between two lines.

To input vectors with the same length into columns, you can use READ command. If you type

READ C1-C3
1 2 3
4 5 6
7 8 9

Using the command PRIN C1-C3, you will see 1 4 7 are in C1, 2 5 8 in C2 and 3 6 9 are in C3.

4.5 Checking the Data Entered and Changing Data

It is always a good idea to check that the program has stored your data properly with the command NAME or PRIN. Typing

NAME

produces a display of the column identifiers together with each column's length, number of missing values, maximum and minimum. Typing

PRINt C...C

where C...C indicates the names or numbers of column of interest. The actual values in these columns can be viewed together with the row numbers and column lengths.

You can use the EDIT command to change a value in a column. The general form of the EDIT command is

EDIT element M in C to read N

Supposing that you discover an error in the 20th case's value of 'AGE', you could input the right value, say, 6.20 by typing

EDIT 20 'AGE' 6.20

You can change all the instances of a particular value into a different value with the CHANge command:

CHANge N in C into M and put results into C

or you can change all the instances in a given range into a new value by typing

CHAN all values between N and N in C to N results to C

5 The Utility Commands

The functions of the utility commands are: selecting cases; sorting cases; splitting columns; joining columns; managing a worksheet; saving and retrieving worksheets; saving operations and results to a text file; writing data columns; and using command files.

5.1 Reorganizing and Managing the Worksheet

5.1.1 Selecting Cases and Dealing with Missing Data

The CHOOse command can be used to select cases. The syntax is

CHOO cases with values of N in C carrying C...C results to C and C...C

or

CHOO cases between N and N in C carrying C...C results to C and C...C

The OMIT command has identical syntax and works in a complementary way, with the indicated rows of data being omitted instead of being chosen. An important use of the OMIT command is to remove cases with unknown values. Suppose the measurement values are in column C2 with the ages in C1, type

OMIT K100 C1 C2 C1 C2
OMIT K99 C2 C1 C2 C1

where K100 or K99 is a indicator of missing value. If -1 indicates the missing value, you can also type

OMIT -1 C1 C2 C1 C2
OMIT -1 C2 C1 C2 C1

5.1.2 Sorting Cases

The SORT command can sort the values in the specified columns into ascending order.

In its basic form, the SORT instruction reorders values in a single column into ascending order.

SORT the values in C putting the results into C

The second form is

SORT using C carrying C...C output into C and C...C

The rows in the carried set of columns C...C are reordered along with the values in the first column C. The number of the output columns is equal to the number of carried columns plus 1.

The third form is used to sort data by the ascending order of two columns. The syntax is

SORT using C C carrying C...C output into C C and C...C 2

Suppose C1 contains codes of subgroups, C2 contains ages, C3 and C4 contain growth measurements of children. Typing

SORT C1 C2 C3-C4 C5 C6 C7-C8 2

The rows in the carried set of columns C3-C4 are reordered along with the ages of C2 in ascending order within the ascending order of subgroups of C1. The number of the output columns are equal to the number of the carrying set of C...C plus 2.

5.1.3 Splitting Columns

The SPLIt command separates a column of values into subcolumns using codes in an identifier column. Suppose you have measurements in C1 and group codes 7, 1, 2, 5 in C2. You can subdivide C1 into C3-C6. Cases whose C2 value is the lowest (1, here) have their C1 values put into C3, and cases whose C2 value is the highest (7, here) have their C1 values put into C6.

SPLI C1 using codes in C2 results to C3-C6

Sometimes you would like to split a measurement column C1 by subranges of a continuous variable of age in C2. First, the GROUp command can be used to change a continuous variable into an ordinal categorical variable. Type

GROUp values in C2 lower limits in C3 result to C4

The group limits must be in ascending order. Values strictly below the first limit give the result 1, those of the remainder strictly below the second limit the value 2, and so on. Then you split the measurement column C1 by C4 with SPLIt command.

5.1.4 Joining Columns

You can join two columns into a single column using the instruction

JOIN values in C and C to form C

The numbers in the second input column are placed after those in the first.

More generally you can specify a string of boxes and columns to be joined, for example

JOIN C1 C4 K1 C10 to be C11

5.1.5 Managing a Worksheet

The command TIDY can recover all unoccupied space in the worksheet and give the message of space available. As the worksheet size is limited, you may run out of space in a lengthy worksheet with a large dataset. If space is tight, it is useful to get rid of unwanted columns by typing

ERAS C...C
TIDY

You could make an entirely fresh start by typing WIPE.

5.2 Using Auxiliary Files

This subsection will describe: saving and retrieving worksheets; saving operations and results to a text file; writing data columns; using command files.

5.2.1 Saving and Retrieving Worksheets

To save the current contents of the worksheet during a worksession, you can type

SAVE

The program prompts for the name of a worksheet storage file and you may give the name of a new file or overwrite an existing worksheet file.

By typing

RETR

the contents of a worksheet file can be retrieved at the start of a worksheet or during a worksession.

5.2.2 Saving Operations and Results to a Text File

You may wish to store your screen output in a disk file which you can edit after exiting from GROSTAT. When you type

LOGOn

for the first time in a worksession, GROSTAT prompts for the name of a log file. Typing

LOGOff

later will suspend logging but the log file you opened with LOGOn remains open. If the system crashes, you'll have lost all the contents you have saved.

5.2.3 Writing Data Columns

To write some of the columns from the worksheet onto a disk file, type

DWRIte C...C

The program will request the name of the file. If more than 10 columns are included, each row of the worksheet will be split into more than one record. If the file is to be read back by GROSTAT, FORTRAN-style formatted input should be used with the format (10G12.5). Missing values and unused spaces on the worksheet will be output as -9.9999E+29.

5.2.4 Using Command Files

If a set of instructions will be used on several occasions and retyping will cost much of your time, you can use an editor to create a command file, typing the commands as if you were using the program interatively. You are restricted to the commands that do not prompt for information. The instruction

ENDObey

must appear on the last line of your file.

To use the command file during a worksession, type

OBEY

You are requested to give the name of the command file.

6 The Graphing Commands

The graphing commands enable you to produce: stem-and-leaf plots, boxplots, basic scatterplots, high-resolution barcharts, and high-resolution plots of points and lines. You can edit, save, and retrieve high-resolution graphs; saving and retrieving a graph.

6.1 Stem-and-Leaf Plots

The instruction of stem-and-leaf plots is

HISTogram of C [using interval width N] [midpoints to C, counts to C]

The program will choose the width of the histogram bars to suit the range of the data. The first option N allows you to set the width of the intervals, and the second gives you a numerical summary of the plot. The midpoints and frequency values can be used in obtaining high-resolution barcharts.

6.2 Boxplots

The boxplot of a sample has a central rectangle bounded by the quartiles, with the median marked by a star. Outside the box, dotted lines extended to readings which are less than a box length outside the quartiles. Values that are between one and one and a half box length from a quartile are marked as a +, and more extreme values are recorded as X's. By typing

BOXPlots of data in C...C

6.3 Basic Scatterplots

The simple form of the instruction

PLOT y values in C against x values in C

will produce a scatter plot, in which a single value is marked as a *'s; 2 to 9 coincident values are plotted as a digit and more than 9 of them are plotted as #.

The other forms allow you to specify the ranges for y and x values:

PLOT values in C between N and N vs values in C between N and N

The command MPLO will plot several Y columns against a single X column by

MPLOt y values in C...C against x values in C

Single points from the successive Y columns are plotted as A's, B's, etc. The number of Y columns must not exceed 26.

The LPLOt command will produce a multiple plot using codes in a column. For example, given columns of weight and height and codes for sex, we can plot weight against height for men and women separately on the same graph. Code values 1, 2, 3,...(less than 26) are plotted as A,B,C.... The instruction is

LPLOt values in C against values in C using codes in C

The scatterplots produced by the commands PLOT, MPLO, or LPOL can be saved in a text file with the command LOGO.

6.4 High-resolution Plots of Points and Lines

A high-resolution scatterplot can be stored internally by typing

PLTP y values in C against x values in C

To see the actual plot on the screen you must type

DISPlay

To leave the display mode at this stage, type Q and press the RETURN key.

You can plot more than one column of Y values against a single column of Xs by using a more general form of the PLTP command

PLTP y values in C...C against x values in C [with symbols N...N]

The value 1 through 6 may be used for N in the optional part of the command.

The high-resolution plotting commands are cumulative permitting superposition of plots. To start afresh you can type

NEWGraph

at any stage.

You can connect the points with straight lines by using the PLTLines command. Type

PLTL using y values in C and x values in C

Note that you may need to sort on the X column first so that the X-coordinates for successive points will be in ascending order.

The PLTLines commands can also be given more than two columns

PLTL using y values in C...C against x values in C [line types in N...N]

By default the Y-columns will use line types 1 to 6 in rotation.

Suppose that the graph you have obtained including some subgroups, say, subdistricts. If you want to print the graphs for these subgroups, type

PLTG codes of subgroups in C y values in C...C x values in C

Note that within subgroups, cases must be sorted by their x values (say, ages), and the codes of the subgroups must be in ascending order before the command PLTG being used.

6.5 High-Resolution Barcharts

The basic form of the command for plotting a barchart is

PLTBars using bar lengths in C and x values in C

As usual you can specify more than one Y column, type

PLTBars for ys in C...C xs in C [shading styles N...N]

Histograms can be 'transformed' into barcharts by specifying to PLTB the columns containing interval midpoints and bar lengths obtained with the HIST command.

6.6 Editing High-Resolution Graphs

When a graph has been displayed, it can be altered interatively using a set of menu-driven subcommands. You can select menu options simply by typing the initial letter (the capital one) and pressing the ENTER/RETURN key.

The main menu appears at the top of the screen when you type the DISPlay Command:

E (dit, U (Se cursor, C (onfigure, I (dentify, H (ard copy, Q (uit

The Edit, Use cursor, Configure, and Identify options each lead you to one or more submenus. Several editing operations use intersecting horizontal and vertical lines to indicate a particular point on the screen. These cursor lines can be moved around the screen using the arrow keys. Movements can be in jumps or steps. A jump is equal to 20 steps and a (default) step is 0.005 of the width or height of the display. The step size can be altered by C option of the main menu.

6.6.1 Printing a graph

H (ard copy, option is shown in the main menu. Type H and press ENTER/RETURN to produce a hard copy of the graph on a printer.

6.6.2 Adding Titles and Text to a Graph

The E (dit option on the main menu leads to a submenu that provides access to the Title subcommand.

T (itles, A (dd text, B (low up, S (cale, E (rase, Q (uit

The T option has its own submenu:

X (title, Y (title, G (title, Q (uit

Each of the title options will prompt you to enter some text and this will appear on the display in the appropriate place when you type RETURN.

The A option on the editing submenu enables you to label points on the screen. It has a submenu

A (dd, Q (uit

and brings up the cursor which can then be moved to a point on the screen. Whenever you select A you will be prompted for some text, and this will be added to the display to the right of the cursor position when you type RETURN.

6.6.3 Enlarging a Portion of a Graph

You can enlarge a rectangular section of a graph to fill the whole screen using the Blow up subcommand. Choosing E from the main menu then B from the editing submenu presents the following submenu

P (oint, Q (uit

and brings up the cursor. You define the rectangle you wish to magnify by marking two points on the screen-the end points of one of the rectangle's diagonals. Move the cursor to one of the points, type P, then press RETURN key. Do the same for the other point. The Q option allows you to return directly to the editing submenu.

6.6.4 Rescaling a Graph/ Altering the Axes

Initially, the graph is scaled automatically. You can make changes on either or both of the axes by selecting the E option of the main menu then the S option of the editing submenu. This gives the submenu

X (scale, Y (scale, Q (uit

Selecting either X or Y gives a further submenu

B (egin, L (ength, G (ap, N (umber interval, F (ormat, Q (uit

call the tick submenu.

An axis consists of a line marked with a number of ticks, some of which have numerical labels. The B, G, and L options each prompt for a number which specifies respectively

* the value at the Begining of the axis;
* the Gap between successive ticks;
* the Length of the axis in terms of the number of ticks.

The N option allows you to specify the number of labelled ticks. If you reply to the N prompt by typing 4, say, every 4th tick will be labelled.

The F option of the tick submenu allows you to change the format of the numeric labels on the axes. It has a submenu

B (efore dp, A (fter dp, E (xponent, Q (uit

The dp stands for decimal point. If you rescale a graph, and then quit and redisplay it later on by typing

DISP 1

You will produce a graph in which the altered scaling is used. If you just type

DISP

the graph will be autoscaled and the rescaling information will be lost.

6.6.5 Altering Cursor Operation

The C option in the main menu produces the configuration submenu

T (rack, O (pen file, H (orizontal, V (ertical, Q (uit

Of these options, H and V allow you to adjust the horizontal and vertical cursor step sizes, respectively. Typing H produces a message asking for the size of horizontal step as fraction of screen width. (The message for V is similar.) In each case a reply of a number n will give a step size of 1/n of the display size.

The T option of the configuration submenu enables you to take measurements from a display on the screen. Selecting it causes the coordinates of the cursor point to appear at the side of the display. It is a toggle; repeating it turns the option off.

6.6.6 Recording Graph Coordinates

You can copy the coordinates of the points you mark on a graph to be recorded into a disk file. Selecting the O option on the configuration submenu will ask you for the name of a file. When this has been given you should return to the main menu by typing Q. If you now select the U option, this will produce a small submenu

P (oint, Q (uit

The cursor will appear and can be moved around the screen. Whenever you select P, the current cursor coordinates (which will appear at the side of the display if you have selected the T option) are added to the chosen file. Typing Q returns you to the main menu and also closes the file.

6.6.7 Identifying Points on a Graph

It is often useful to identify a plotted point by finding out which row on the worksheet its coordinates came from. To do this after the graph has been displayed, selected the I option from the main menu. This produces the submenu

P (oint, Q (uit

and brings up the cursor. If you now set the cursor on the required point and select P, the column of the X and Y coordinates will be displayed along with the serial number of the point in these columns.

6.7 Saving and Retrieving a High-Resolution Graph

The commands are GSAVe and GRETrieve, both of which will prompt you for a file name. Note that the data for a graph which has been RETRieved will overwrite any existing data in the columns on the worksheet that were originally used to make the graph.

7 The Basic Arithmetic and Statistics Commands

This section describes the arithmetic instructions and the commands for computing descriptive statistics, tabulating data, obtaining correlations and tail probabilities.

7.1 The Algorithm Commands

The one-operand instructions and their functions are as follows

LOGTen C C	log to the base 10
LOGE C C	log to the base e
LOGIt C C	ln(p / (1-p)) for $0<p<1$
ROUNd C C	round to the nearest integer

NED C C	normal equivalent deviate of q, for 0<q<1
SIN C C	sine of an angle in radians
COS C C	cosine of an angle in radians

The one-operand instructions can work with boxes as well as columns. For example,

LOGT C C

and

LOGT K K

are the general forms.

The two-operand instructions specify two inputs and a destination for the output. The commands have several general forms, as illustrated here with the ADD command:

ADD C C C
ADD C N C
ADD N C C

and

ADD N N K

Recall that N can represent a number or a box. When both inputs are columns, the commands work row-wise.

The list of these is

ADD	first to second
SUBTract	first from second
MULTiply	first by second
DIVIde	first by second

7.2 Descriptive Statistics

The AVERage command can produce the number of the values in a column, their mean, the standard deviation and the standard error of the mean. The syntax is

AVERage values in C

If you provide box indicators, all four quantities will be stored. The general form of the instruction is

AVERage values in C [results to K K K K]

It is possible to get the averages and standard deviations of a whole set of columns by typing

AVER N columns C...C [means to C and s devs to C]

The MOMEnts command calculates the first four moments of the data in the indicated column. The instruction is

MOMEnts of data in C [store in K K K K]

Actually what are calculated are the k-statistics, estimates of the cumulants (Kendall & Stuart, vol.1, section 12.6). Also given are Fisher's measures of skewness and kurtosis g1 and g2 together with their standard errors. The k-statistics will be stored if boxes are provided.

7.3 Tabulation

The program can form two sorts of table-frequency tables, where each cell contains simply a count of the number of cases failing in the cell, and means tables in which each cell contains, in addition to the count, the mean and standard deviation of a particular variable. Each kind of table can have either one or two factors classifying the cases.

The general forms of the command for creating a frequency table are as follow:

TABUlate with output mode N factor levels in C
TABUlate mode N factor levels in C and C

The output mode is a number or box which calls for extra output in addition to the simple counts. The number is built up from several one-digit codes. Three of these call for percentages to be printed:

* 1 indicates percentages of row totals
* 2 indicates percentages of column totals
* 3 indicates percentages of the grand total

(these are all equivalent for one-way tables). Multiple codes are simple placed side by side. Thus an indicator 13 (31) calls for row percentages and total percentages to be printed. If no extra output is needed, the indicator should be zero.

With a two-way table, a further code equal to 4 produced χ contributions. These are generalized residuals of the form (observed count - expected count)/(expected count) and are such that their sum of squares is the usual χ^2 value for the table. (This value is also printed). You can get the value of the χ^2 statistic for the test of the hypothesis of independence of the factors in a multi-way table along with the significance probability by typing

CHISquared for table in columns C..C

If the table is 2 x 2, a continuity correction will be used.

The instruction for a one- or two-way means table is very similar:

TABUlate means and sds of values in C factors in C [and C]

When a means table is printed, the standard deviations in the margins of the table are obtained by pooling within the appropriate cells of the table. Note that when a two-way table is output, the first factor mentioned forms the columns, the second factor the rows.

Any two-way table can be restricted to entries with a specified value (or with values between specified limits) of a third values. The general form of the command are:

TABUlate (N or C) factor C and C with C equal to N
TABUlate (N or C) factor C and C with C between N and N

7.4 Correlations

If we have two columns of data, the instruction

CORRelate values in C and in C [result to K K K K K]

will produce the number of values, both means, both standard deviations and the correlation coefficient with its significance probability. All these (except the last) will be stored if box indicators are provided. A complete correlation matrix of the data in two or more columns can be obtained by the instruction

CORMatrix of N variates in C...C and store it in C

The means and standard deviations are also displayed. Note that a row of data containing one or more unknown values is omitted. If an extra column is specified, the correlation coefficients (omitting the 1.0's on the diagonal) will be stored in the sequence 12 13 23 14 24 34 15...

7.5 Tail Probabilities

The commands described here are for the tail probabilities of the standard Gaussian, t, χ^2, and F distributions, respectively using

NPROability for value N
TPROability for value N with N df
CPROability for value N with N df
FPROability for value N with N and N df

These all give upper-tail probabilities. The probabilities will be stored if a box indicator is provided.

7.6 Multiple Regression

You can do regression calculations, either simple or multiple, by using the instruction of the form

REGR responses in C on N predictors in C...C [yhats to C coeffs to C]

The value of N, the number of explanatory variables, must not exceed 20. The output from this instruction includes the analysis of variance table and a list of the regression coefficients and their standard errors. By specifying two extra columns in the instruction, it is possible to store the fitted values from the regression and also the coefficients (with the intercept term coming last).

It is sometimes necessary to do regression analysis with no constant term-regression through the origin, or regression with zero intercept. This can be done by replace the REGRession keyword by OREG.

8 The Growth Curve Fitting Commands

This section describes the commands for growth centile curve fitting. The commands are for obtaining raw centiles, fitting ordinary polynomials, fitting piecewise polynomials with constraints and fitting smoothly joined piecewise polynomials.

8.1 The command CBOX for Obtaining Raw Centiles

The method of obtaining raw centiles is fully described in Healy and Rasbash(1987). Briefly, the data points should be imagined as plotted in a scatter diagram. A 'box' is superimposed on the left-hand end of the diagram so as to contain a fixed proportion of the data points. A regression line is fitted to the points in the box, the residuals are calculated and sorted, and interpolated centiles are plotted against the median of the x-values in the box. The box is then moved to the right by one point or some points (if 'jump size' is specified) and the process is repeated, moving the box across the diagram until all the data have been covered. In this way a set of 'raw' centils curves are constructed. The instruction for obtaining initial centiles is

CBOX using C C p C output C...C [using N] [using C output C] where the first two columns are the y and x data columns, holding the measurement and age values respectively. Next comes a number p specifying the size of the box as a fraction of the total data. The next column holds the percentages defining the required centiles in ascending order. Last comes a set of columns C...C which will hold the raw centile points. If m centiles are to be calculated, there should be m+1 columns specified with the last one holding the median x-values. In the optional part, N is the number of points between consecutive box positions (specified as 'jump size') that is, the number of points omitted and new points entered into the box each time the box moves along the age axis; the first C contains the codes of subgroups and the second C is the output column for codes of subgroups corresponding to the centile estimates.

Note that before the command CBOX is used, the measurement data must be sorted on age, or age within subgroup if centiles for subgroups are being estimated.

A jump size of three points reduces the number of points in the raw centiles by a factor of three and thereby speeds up the time taken to smooth them threefold. If you experiment with this parameter you will often find that centile estimates produced with different jump sizes are almost identical when large samples are available.

The choice of box size is important. As the box size increases, fluctuations in the raw centiles become less marked. If you make the box too large, however, real effects in the data may be smoothed over. For example nuances of growth during the adolescent spurt may be lost. You should experiment with the box size.

The magnitude of the most extreme centiles you wish to estimate is a crucial factor in determining the appropriate box size. The more extreme they are the larger the box must be. In terms of number of points, the minimum box size that the program will accept is given by :

$$N = \mathrm{Max}\left\{\frac{0.5}{P_l}, \frac{0.5}{(1-P_h)}\right\}$$

where P_l is the proportion corresponding to the lowest centile to be estimated and P_h is the proportion corresponding to the highest centile to be estimated. Experience suggests that a reasonable box size is between 3 and 4 times the minimum limit.

The age range spanned by the raw centiles produced by small boxes is wider than that by large boxes. This is because the raw centiles begin one half a box width in from the leftmost data point and end one half a box width before the rightmost data point. With small data sets, if you use the above rule to select the box size and try to estimate very extreme centiles you may find that the age range spanned by the raw centiles is unacceptably narrow. In this case you will probably have to estimate less extreme centiles.

Some examples of experimenting with box size will be shown in section 9.1.1

8.2 Commands for Fitting Curves

This subsection describes the commands for fitting polynomials with ordinary polynomials and piecewise polynomials with smooth joins.

8.2.1 The command CPOL for Fitting Ordinary Polynomials

Fitting centile curves with ordinary polynomials is described by Healy and Rasbash (1987). The model will be reviewed in brief.

First it is assumed that the curve for each of the centiles can be fitted by a polynomial of degree p. If t denotes age and y_i the smoothed values of the ith centile, we have

$$y_i = a_{0,i} + a_{1,i}t + a_{2,i}t^2 + \ldots + a_{j,i}t^j + \ldots + a_{p,i}t^p \qquad (1)$$

Suppose that P_i is the proportion labelling the ith centile and that z_i is the Normal Equivalent Deviate (NED) of P_i thus for the 50th centile $P_i = 0.50$, $z_i = 0.00$, for the 3rd centile $P_i = 0.03$, $z_i = -1.88$. We then assume that the coefficients for a given j can be expressed as a polynomial in z_i

$$a_{j,t} = b_{j,0} + b_{j,1}z_t + \ldots + b_{j,q_j}z_t^{q_j}$$

(2)

where the degree q_j of the polynomial may differ from one value of j to another and will usually be higher for the low-order coefficients and may be zero for the high-order ones.

The command CPOL is

CPOL using N...N C C...C output to C...C C

The first set of number specify p, q_0, q_1, ..., q_p. Next comes the column holding the m required centiles, that is, the m defining percentages, followed by the $m+1$ columns holding the raw centile data (including the ages). The next set of columns must be m in number and will hold the smoothed centile values. The estimated polynomial coefficients are output to the final column. First will come the coefficients of z for the intercepts in ascending order (constant, linear term, quadratic term,...), next will be the coefficients of zz for the linear term in age, and so on.

8.2.2 The Command PCWS for Fitting Piecewise Polynomials

The method is fully described by Goldstein and Pan (1990), which constructs two or more polynomials joined in a smooth fashion at specified joint points. The procedure is flexible and efficient in studies of child growth when estimates are required over a wide age range, where a single polynomial is inadequate. The method allows estimation for all subgroups simultaneously, e.g., for males and females. It is assumed that subgroups differ only in the intercept and linear coefficients.

The syntax is

PCWS using C C...C C C...C C output to C...C C using [N] [C] [C] [C]

Suppose 5 centile curves are to be estimated with two joint points. Type, for example,

PCWS C11 C12-C14 C50 C1-C5 C6 C21-C25 C26

where C11 contains 2 values which are the joint points of age in ascending order. C12-C14 are columns in each of which we have the value of p, the order of the polynomial, and q_0-q_p, the parameters of the model for Z. C12 is for the first subrange, C13 for the second subrange and C14 the third. C50 contains the required centiles, say, 3, 25, 50, 75 and 97. C1-C5 contain the unsmoothed centiles and C6 contains their age values produced by the CBOX command. Smoothed centiles are output to columns C21-C25. (NOTE-no column of x need be specified for this last group of output columns.) The coefficients of the smoothed polynomials are output to C26 for subranges of each subgroup (from the first subgroup to the last one). The q_0 coefficient of the polynomial fitted to the constant term is output first followed by the q_1 coefficient and so on for each subrange.

In the optional part, the first N is a value from 1 to 7. This number indicates how many and which of the optional columns will follow. The subgroups codes column would procede the centiles weighte column, which in term would precede the ages weights column.

N	codes for subgroups	weights for centiles	weights for ages
1	yes	no	no
2	yes	yes	no
3	yes	no	yes
4	yes	yes	yes

5	no	yes	no
6	no	no	yes
7	no	yes	yes

If you are using centile weights, the column containing them must contain one entry for each centile. If you are using age weights the column containing them must contain one entry for each age point.

Note that the raw centiles must be sorted on age, but NOT on age within subgroup before the command PCWS is used. It means that if raw centiles are obtained by CBOX using the optional subgroup column, they must be sorted again only by age before this command is used.

Limitation: The maximum numbers of subranges and subgroups are 4 and 10 respectively. No more than 60 coefficients may be used.

Sometimes you may fail to obtain smooth centile curves and you will need to experiment with chosing orders of polynomials. Here are some suggestions describing what orders of p, q_0, q_1,...should be fitted across the centile distribution in neighbouring age ranges if the curves for each range are to join smoothly. The suggestions are:

For odd numbered ranges(the first, third, fifth etc) -

> ***The order of polynomial fitted across the centile distribution to the intercept and at least one other of the coefficients of the age polynomial, must be one less than the total number of centiles being estimated.***

For even numbered ranges you may specify any model you like.

8.2.3 The Command SPLN for Fitting Piecewise Polynomials

You can also fit smoothly joining piecewise polynomials using 'grafted polynomials' to obtain smoothly joining centile curves.

It is assumed that all the curves can be fitted by a polynomial of degree p. If t denotes age and Y_i the smoothed value of ith centile, we have

$$Y_i = a_{0,i} + a_{1,i}t + \ldots + a_{p,i}t^p + a_{p+1,i}(t-c_1)_+^p + \ldots + a_{p+m-1,i}(t-c_{m-1})_+^p \quad (1)$$

where

$$(t-c_l)_+^p = (t-c_l)^p \quad \text{when} \quad t > c_l$$
$$0 \quad \text{when} \quad t \le c_l \quad l = 1,2,\ldots,m-1$$

c_l is the lth joint point between the age groups and $c_1 < c_2 < \ldots < c_{m-1}$. Of course, when $m = 1$, the new method is the same as that presented by Healy et al. For example, with one joint point and polynomials of degree 3 we obtain

$$Y_i = a_{0,i} + a_{1,i}t + a_{2,i}t^2 + a_{3,i}t^3 \qquad \textit{when} \quad t \le c_1$$
$$= a_{0,i} + a_{1,i}t + a_{2,i}t^2 + a_{3,i}t^3 + a_{4,i}(t-c_1)^3 \qquad \textit{when} \quad t > c_1$$

It is clear that the two components of the curve join smoothly at the joint point, and this is generally true for curves of the form (1). The value of m and c_l can be chosen after inspection of the data using existing knowledge.

The coefficients in (1) are now regressed on a polynomial function of the NED's Z_i of degree q.

where the Z_i, the degree p, and the degree q_j of the polynomial hold the same meaning as in 8.2.1.

$$a_{j,i} = b_{j,0} + b_{j,1}Z_i + \ldots + b_{j,q_j}Z_i^{q_j} + e_{j,i} \qquad j = 0, 1, \ldots p + m - 1$$

(2)

In general the value q_j will usually be higher for the low-order coefficients and may be zero for high-order ones. Some exploration will generally be needed to obtain optimum values for p and q_j.

The instruction SPLN has the form

SPLN using N...N C C C...C C output C...C C [C]

Suppose 5 centile curves are to be estimated with two joint points, type

SPLN 3 2 1 1 0 C11 C50 C1-C5 C6 C21-C25 C26

The first set of number 3 2 1 1 0 specify p, q_0, q_1, ..., q_p respectively. C11 contains the joint points in ascending order. C50 contains the required centiles, say, 3, 25, 50, 75, 97. The columns C1-C5 contain unsmoothed centiles and C6 age values related to the unsmoothed centiles produced by the CBOX command. Smoothed centiles are output to C21-C25. The estimated coefficients are output to C26.

The optional column contains the codes for subgroups.

Note that the raw centiles must be sorted on age, but NOT on age within subgroup before command SPLN being used. This means that if raw centiles are obtained by CBOX with subgroup column, they must be sorted again only by age before command SPLN is used.

Limitation: The maximum numbers of subranges and subgroups are 4 and 10 respectively. No more than 60 coefficients may be used. Sometimes the procedure may fail to fit the curve properly at adulthood where growth ceases. In this case it may be possible to use the PCWS command.

8.3 Commands for Checking Goodness of Fit

There are several ways to check the goodness of fit. Visually, you can plot the smoothed centile curves with the unsmoothed centile points superimposed or plot the smoothed centile curves with measurement data superimposed. A test of fit can be obtained by counting the number of data points that fall between the centile curves for a box in the way done with CBOX command and comparing these counts with their expected values. The counting can be obtained by the command CONT with the coefficients estimated by the commands CPOL or PCWS but not by SPLN. The CONT instruction is

CONT using C C p C C C...C C output C...C C [using N] [using C]

Suppose we use CONT to check goodness of fit of the above PCWS procedure, that is, 5 smoothed centiles with two joint points. Type,

CONT C8 C9 0.1 C50 C11 C12-C14 C26 C31-C35 C36

where C8 contains the original measurements. C9 contains age values. 0.1 is the box size as a proportion of the dataset. C50 contains the requested centiles, say, 3, 25, 50, 75, 97. C11 contains the joint points in ascending order. C12-C14 contain the values p, $q_0 \ldots q_p$ for Zs in each subrange. C26 contains coefficients obtained by PCWS. The proportion in each centile group are output into C31-C35 and related ages into C36.

In the optional part: N is jump size, and C is the column containing the codes of the subgroups.

The selecting of the box size and jump size is important and will be illustrated in section 9.1.2.

Note that the measurement data must be sorted by age NOT by age within subgroup before the command CONT is used. This means that if raw centiles are obtained by CBOX using the subgroup column, they must be sorted again only by age before this command is used.

The command CVAL is also available to check the goodness of fit for CPOL and PCWS.

CVAL using C C C C...C C output C {using C }

Suppose we use the CVAL to check goodness of fit for the above PCWS procedure with two joint points. Type

CVAL C8 C9 C11 C12-C14 C26 C30

where C8 contains the original measurements. C9 contains age values. C11 contains joint points. C11-C14 contain values $p, q_0 \ldots q_p$ for Zs in each subgroup. C26 contains coefficients obtained by PCWS. The values of the proportions in each centile group are output into C30.

In the optional part is the column C which contains the codes for subgroups.

Note that you must be sure that the raw centiles are sorted by age NOT by age within subgroup before the command CVAL is used. This means that if raw centiles are obtained by CBOX with the subgroup column, they must be sorted again by age before this command is used.

9 Examples of Growth Curve Fitting

The file weight.dat contains three variables: age, weight, and sex, respectively. It includes 425 boys and 442 girls aged from birth to 72 months, and it is a cross-sectional dataset. The codes for sex, that is, subgroups are 1 for males and 2 for females. Read the data into columns 1-3 of the worksheet and save the worksheet by

DREAd C1 C2 C3
WEIGHT.DAT
SAVE
WEIGHT.WS

This section details a number of different smoothing models of increasing complexity.

9.1 Models with Only One Group

This subsection will show the estimation of centile curves with the commands CPOL, PCWS and SPLN when there is only one group, and checking the goodness of fit for these models. The example used in this subsection uses only males. You can choose the data for males from the worksheet you have just produced by typing

CHOO all values of 1 in C3 C1 C2 output to C3 C1 C2 {pick males}

If there are some missing values in your data, you should delete the records of the missing data with the command OMIT. There are no missing values in weight.dat.

9.1.1 Using CBOX to obtain raw centiles

If you type

NAME

you will find there are 3 columns in the worksheet. You can name each column as follows

NAME C1 'AGE' C2 'WEIGHT' C3 'MALE'

Suppose you wish to obtain 5 raw centiles, such as, 3rd, 25th, 50th, 75th and 97th, type

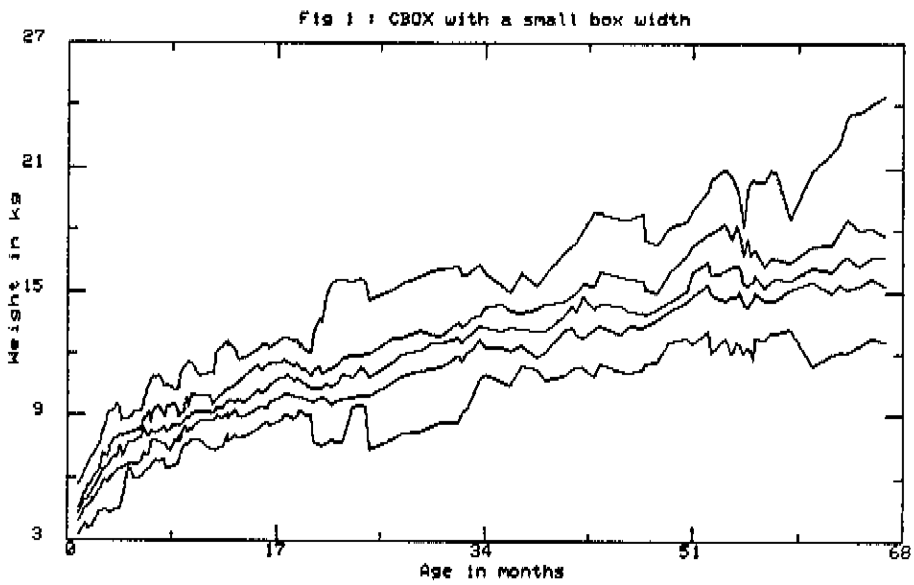
```
SET C50
3 25 50 75 97    {store list of centiles to be estimates in C50]
SORT C1-C3 C1-C3  {sort data into ascending age order}
CBOX C2 C1 0.05 C50 C4-C9 3
```

The last parameter on the CBOX command is optional. It specifies the 'jump size' for the counting box. If this parameter is omitted a jump size of one data point is used. If you experiment with this parameter you will often find that centiles estimates produced with different jump sizes are almost identical when large samples are available. The full syntax of the CBOX command is described in subsection 8.1.

The choice of box size is important. As the box size increases, fluctuations in the raw centiles become less marked. To demonstrate this first plot out the raw centiles you have just estimated by typing

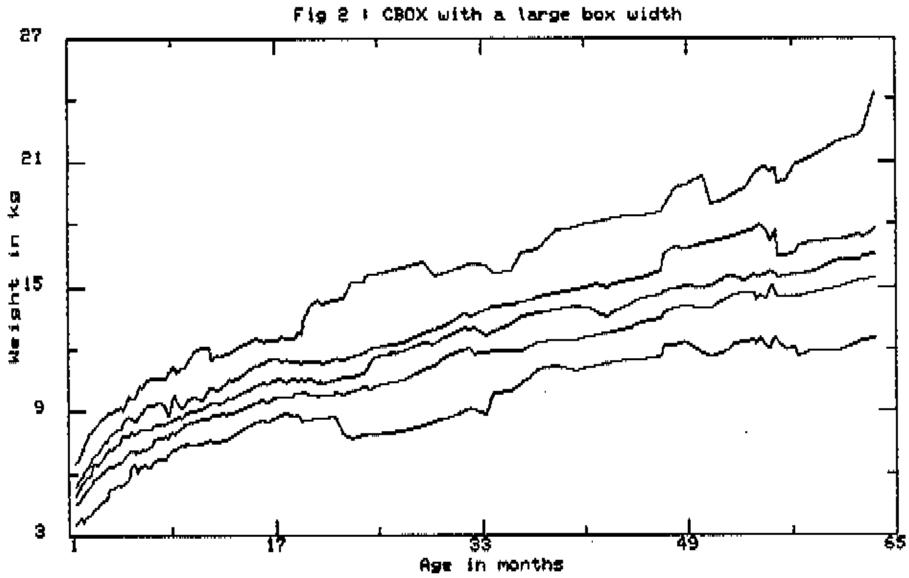
```
PLTL C4-C9 1 1 1 1 1
DISP
```

which will produce the following graph.



Now try doubling the box size and plotting out the resulting raw centiles :

```
CBOX C2 C1 0.1 C50 C4-C9 3
NEWG
PLTL C4-C9 1 1 1 1 1
DISP
```



Studying these two figures it is clear that the raw centiles produced by the larger box size are less jagged. If you make the box too large then real effects in the data may be smoothed over. For example nuances of growth during the adolescent spurt may be lost. You should experiment with the box size. More details have been discussed in subsection 8.1.

9.1.2 Using CPOL to Fit Polynomials

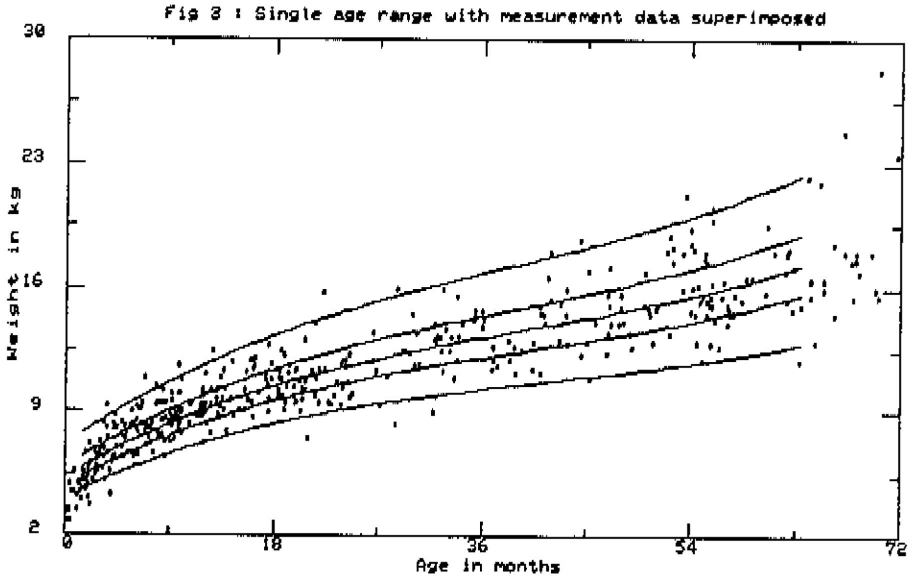
Before moving on we'll attempt a piecewise model in which the age range is split by join points. To begin with, we will try fitting a single age range model with the CPOL command which is also available in GROSTAT I. This will serve two purposes. Firstly, a single age range may provide a satisfactory solution, in which case we need go no further. Secondly, if the resulting curves are unsatisfactory studying them should give some insight as to where to position age ranges in a more complex model.

Studying the raw centile curves indicates that they could be summarized by fitting a cubic polynomial along the age axis and that the centile distribution is very asymmetric and varied markedly along the age axis. The worksheet used here is from section 9.1.1. A suitable model might then be fitted by

```
CPOL 3 2 1 1 1 C50 C4-C9 C14-C19
```

plot out these curves with the original measurements superimposed:

```
NEWG
PLTL C14-C18 C9 1 1 1 1 1
PLTP C2 C1
DISP
```



It is clear from Figure 3 that these centiles fit the data poorly at the lower end of the age range, they should be increasing much more sharply. We could fit higher order polynomials to allow for this. Trying this does produce some improvement, but unfortunately some of the extra flexibility is used up at the higher end of the age range and there is a tendency for the centiles to start decreasing(as children get older) at the higher end of age range (not illustrated here). Prior knowledge of the way children grow tells us such centiles are very improbable. You can experiment further to demonstrate to yourself the limitations of these single age range models.

The command CONT will show you, graphically, the goodness of fit of any given set of centiles. The command needs the orders of the polynomials fitted to be placed into a column. To execute the command type

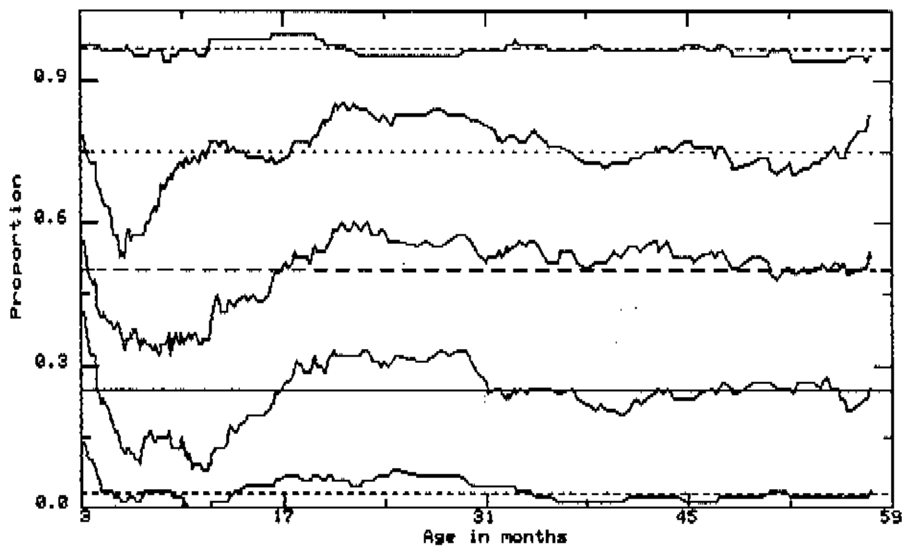
```
SET C51
3 2 1 1 1    {put orders into column 51}
CONT C2 C1 0.1 C50 C80 C51 C19 C24-C29 3
```

The first two columns contain the measurement values and age at measurement. The third number is the box width expressed as a proportion of data set size. The fourth parameter is a column containing the centiles for which counts are desired. The fifth parameter is a column specifying what age ranges were used when estimating the centiles. If a single age range was used just put any empty column here, say, C80. The sixth and seventh parameters contain the polynomial orders of the fitted model and the estimated coefficients. The output is written to the next group of columns; there should be one column for each centile involved plus an extra column for the age values associated with the counts. The last parameter specifies the box jump size. The full syntax of the command is given in section 8.3.

Plot the proportions less than each centile, superimposed on their expectations:

```
NEWG
PLTL C24-C29 1 1 1 1 1
READ C44-C49
0.03 0.25 0.50 0.75 0.97 3
0.03 0.25 0.50 0.75 0.97 59
PLTL C44-C49
DISP
```

Fig 4 : Proportion of points below each centile



At ages where these centile checking curves(solid lines) are below their expectations (dashed lines) the actual centile curves are too low. Conversely where they are above their expectations the actual centile curves are too high. Studying such graphs in conjunction with graphs of the actual centiles with the measurement data superimposed can reveal systematic errors of fit which may need correcting. Note that the variation of the counts about their theoretical values will decrease as the box size gets larger.

It is worth noting that these checking curves are produced by moving a counting box through the smoothed centiles which themselves are produced by a similar process. Therefore at both ends the checking curves are now a full box width short of the entire age span of the data. This means the centiles are not fitted at both ends or they can not be adequately checked in the end regions. One solution to this problem is to collect data over a greater age range than that for which you wish to estimate centiles. Of course this strategy breaks down at one end if you wish to estimate centiles from birth, and it will be necessary to collect large samples at birth and just after birth to obtain accurate estimates there. In general, if you have a large number of measurements at an extreme age then one box worth of points corresponds to a very small distance on the axis, and minimizes the axis length of the regions effected by inadequate end point estimation.

You could derive a figure for an overall goodness of fit, to help you compare models, by taking the sum of the squares of the estimates from their expectations. Do not set too much store by such summarizing statistics because the model which minimizes them may produce unrealistic centiles which are simply following the contours of sampling error. It is far better to use graphical techniques to give a broad idea of goodness of fit combined with knowledge of how children grow in deciding what models are appropriate.

The danger of fitting high order models which produce undulating centiles that merely trace out the idiosyncrasies of a sample has already been described in section 2.2.

9.1.3 Using PCWS to Fit Piecewise Polynomials

As noted earlier the centiles in Figure 3 are poor at the bottom end of the age range. We will divide the age range into two sections 0-12 months and 12-60 months and fit separate smoothing polynomials to each range, constraining them to join smoothly. Hopefully this will provide centile estimates which fit the data better in the first age range. We start with a simple model fitting a cubic to the first age range and a quadratic to the second. The worksheet used here is from section 9.1.1. The commands to set up this model are

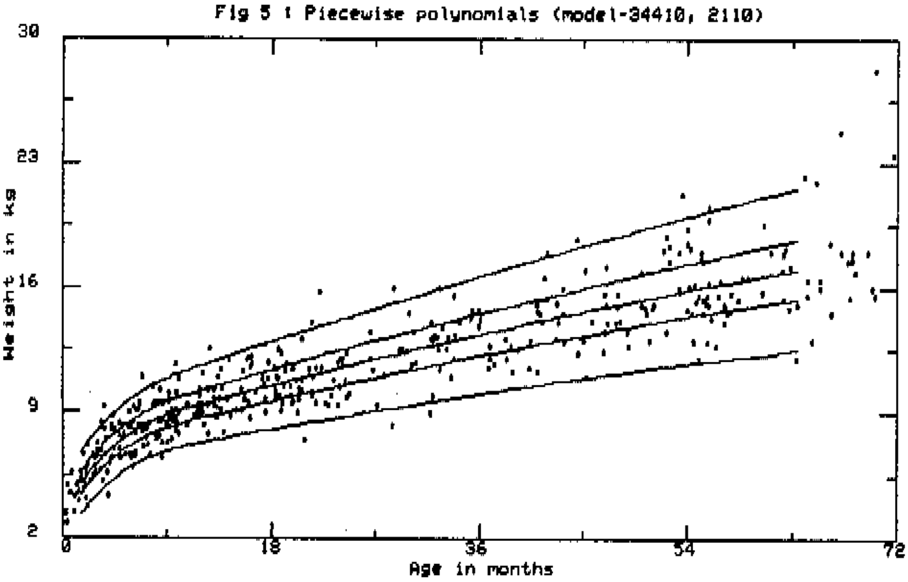
```
SET C51      {put join point of ranges in column C51}
12
SET C52      {put model for the first age range in C52}
3 4 4 1 0
SET C53      {put model for the second age range in C53}
2 1 1 0
PCWS C51 C52 C53 C50 C4-C9 C24-C28 C29   {fit model}
```

The first column of the PCWS contains the join point(s) which define the subranges to be fitted. The next set of columns(C52, C53 in this example) contain the model to be fitted to each range, one column per range. The next column(C50) contains the required centiles. Then come the raw centiles which are to be smoothed and their associated age values(C4-C9). Next, output columns for the smoothed centiles(C24-C28). No column for age values is required here since they have the same ones as the raw centiles. Finally a column(C29) for the estimated coefficients.

Sometimes you may fail to get smoothly joining curves. In that case, you can refer to our suggestions on selecting orders of polynomials (in subsection 8.2.2) with PCWS or you can use the command SPLN to fit piecewise polynomials. To graph the smoothed centiles type:

```
NEWG
PLTL C24-C28 C9 1 1 1 1 1
PLTP C2 C1
DISP
```

Fig 5 : Piecewise polynomials (model-34410, 2110)



This is clearly an improved fit at the lower age range. Try using the B(low up facility, in the graphics E(dit sub-menu to examine the fit at the lower end more closely. Do this for each model and compare results. Type

```
NEWG
PLTL C24-C28 C9 1 1 1 1 1
PLTP C14-C18 C9 1 1 1 1 1
DISP
```

Figure 6 shows the single age range model superimposed on the piecewise model. Unquestionably, the piecewise model allows the curves to rise more sharply at the beginning of the age range. Figure 7 is an enlargement of the area of interest to illustrate this point more clearly.

Fig 6 : Single age range (dots) with piecewise (lines) smoothing

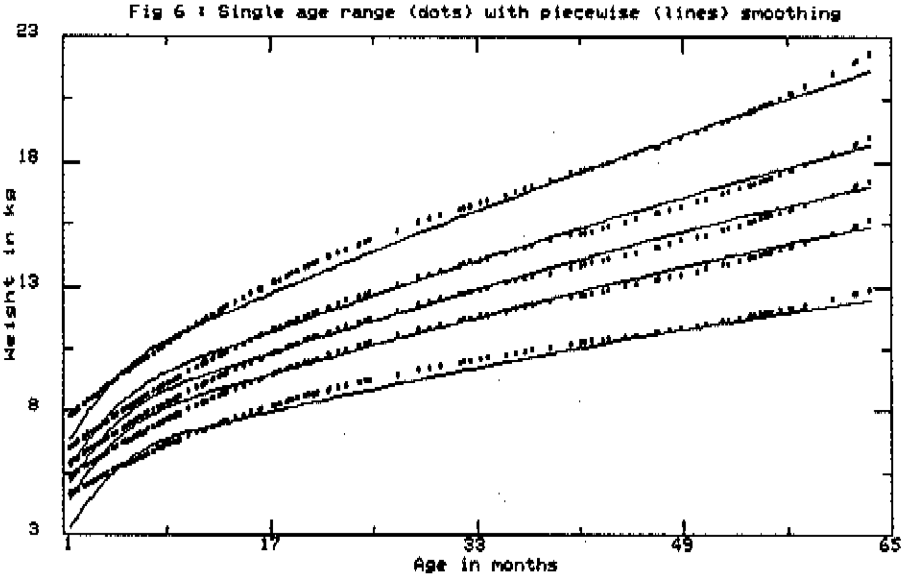
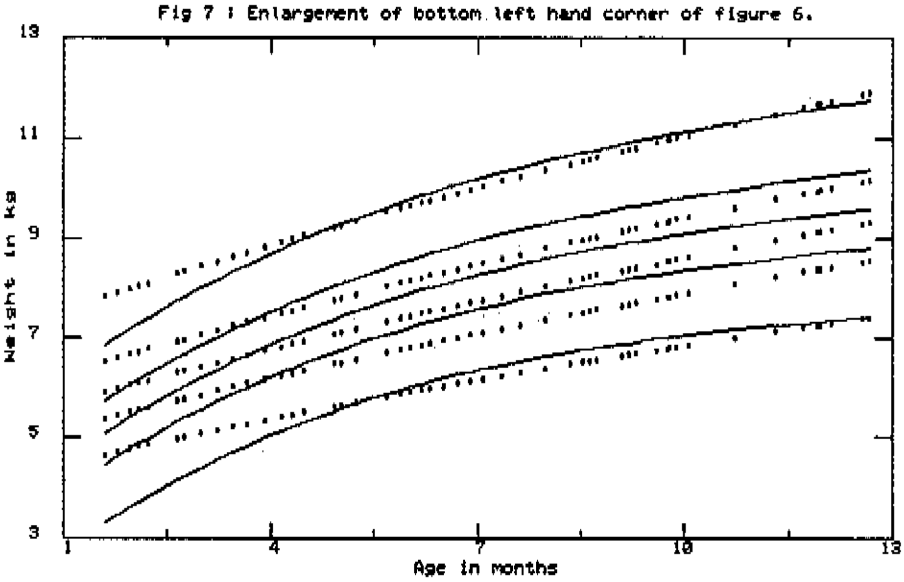
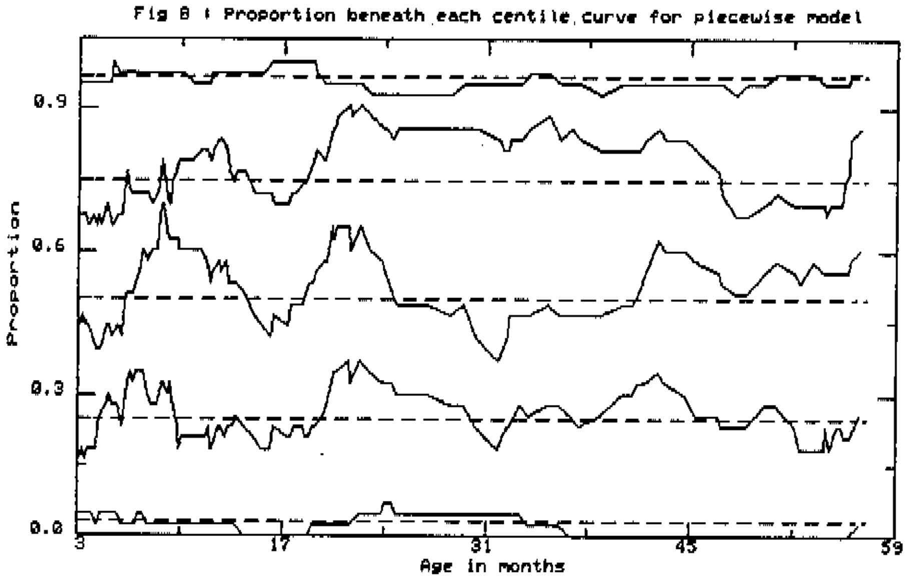


Fig 7 : Enlargement of bottom left hand corner of figure 6.



We can get an idea of an idea of how well these centiles describe this sample by using the CONT command and graphing the results :

CONT C2 C1 0.1 C50-C53 C29 C34-C39 3
NEWG
PLTL C34-C39 1 1 1 1 1
PLTL C44-C49 2 2 2 2 2



Compare this with the equivalent graph for the single age range smoothing model (Figure 4). The pattern is different for the first age range and quite similar for the second. Try fitting other piecewise models to see if you can produce even better fitting curves.

9.1.4 Using SPLN to Fit Piecewise Polynomials

The command SPLN is used to fit piecewise polynomials. With the worksheet from section 9.1.1, you can put join point/points in one column. Suppose 12 months is selected to be the join point to split the whole age range into two subranges. Type

```
SET C51     {input joint point}
12
SPLN 3 2 1 1 0 C51 C50 C4-C9 C22-C27
```

The syntax of command SPLN above is similar to CPOL except for the inclusion of C51 in SPLN. You can plot the smoothed centiles with the raw centiles superimposed by typing

```
NEWG
PLTL C22-C26 C9 1 1 1 1 1
DISP
PLTP C4-C9
DISP
```

The graph is similar to Figure 5.

Also you can select other parameters of p, q_0, q_1,..., and perhaps, join point/points. For example, you may put two join points by typing

```
SET C51
12 36
SPLN 3 2 1 0 0 C51 C50 C4-C9 C22-C27
```

9.2 Models with Subgroups.

Now try fitting the boys and girls simultaneously. The advantage of this approach, as mentioned earlier, is that because there is some commonality between the way boys and girls grow, both sets of curves can be made more precise than if they were estimated separately, sharing this common information.

The dataset weight.dat mentioned at the begin of section 9 will be used here. Read the data from weight.dat into column 1-3 (we omitted the girls' measurements in section 9.1) or retrieve weight.ws with the command RETR.

9.2.1 Using CBOX to Obtain Raw Centiles with Subgroups.

You must reissue the CBOX command specifying the subgroup information. When working with subgroups CBOX requires the data to be sorted by age within group.

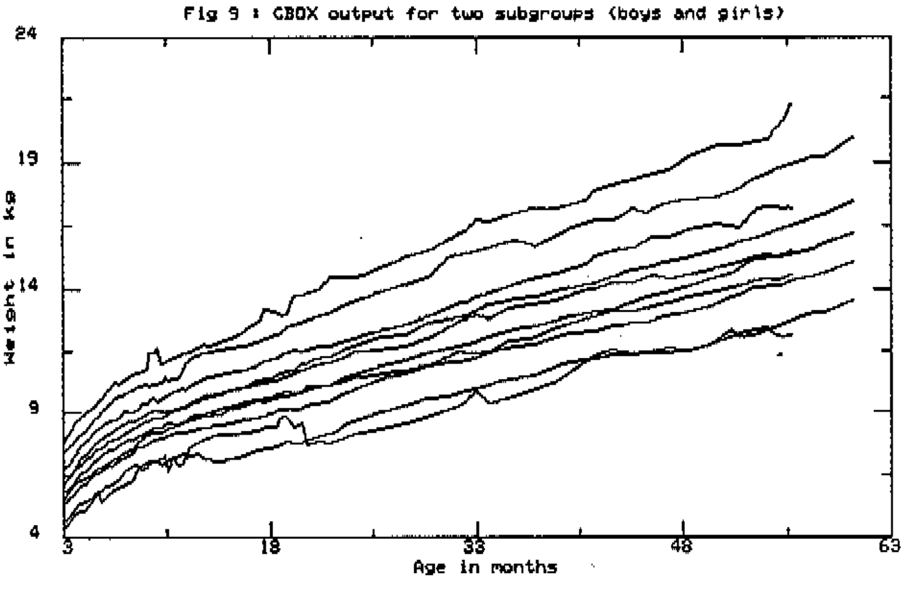
```
RETR
WEIGHT.WS    {the worksheet saved at the begin of section 9}
SORT C3 C1 C2 C3 C1 C2 2  {sorted by age within subgroups}
SET C50
3 25 50 75 97
CBOX C2 C1 0.1 C50 C4-C9 5 C3 C20
```

The final 2 on the SORT command means sort on the first two columns as opposed to the first column only(the default). The CBOX command now has an extra two columns tagged on the end. The first contains the group information and this column must be the same length as the columns holding the measurement data. The last column contains the codes specifying whether each raw centile output point belongs to a boy or a girl.

This command now gives us two sets of raw centiles one for each group. We can plot these out using the PLTGroup command :

```
NEWG
PLTG  groups in C20 data in C4-C9
DISP
```

Fig 9 : CBOX output for two subgroups (boys and girls)



Here you can see the two sets of centiles, the ones for the males spanning a slightly shorter age range.

A quick word on the PLTGroup command. The data to be plotted must be sorted so that all members of a group are together and the x values within a group are in ascending order. In this case we have already sorted on age within group so no re-sorting is required.

9.2.2 Using PCWS to Obtain Piecewise Polynomials with Subgroups.

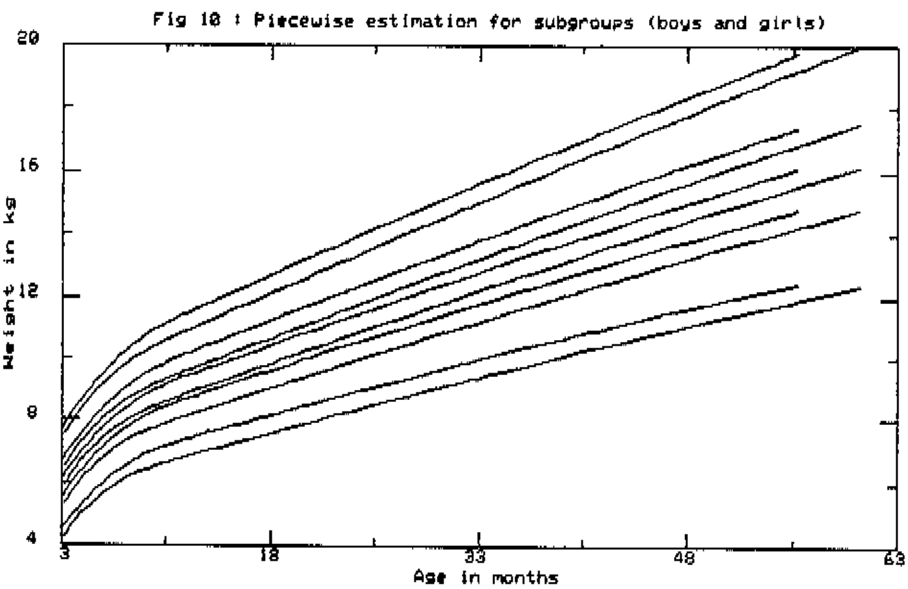
For each subgroup, we first fit the model we ended up with for the boys in section 9.1. The PCWS command requires the data to be sorted in age order. This facilitates fitting around the cut points. At the moment the data are sorted by age within subgroups. To re-sort the data and fit the model :

```
SET C51
12
SET C52
3 4 4 1 0
SET C53
2 1 1 0
SORT C9 C4-C8 C20 C9 C4-C8 C20
PCWS C51-C53  C50 C4-C9 C14-C19 1 C20
```

The syntax for the PCWS command is the same as before with an extra couple of parameters, specifying the subgroup information, added at the end. The last column C20 contains the group codes. There are several extensions to the basic PCWS command (see section 8.2.2). You may fit subranges, weight the centiles, weight individual age points or any combination of all or some of these. Each combination is given a number as in section 8.2.2. The number for "subgroups only" is 1. Hence the penultimate parameter in this case is 1.

To plot out both sets of centiles on one graph using the PLTG command we must re-sort the data : age within subgroup.

```
SORT C20 C9 C14-C18 C20 C9 C14-C18 2
NEWG
PLTG C20 C14-C18 C9
DISP
```



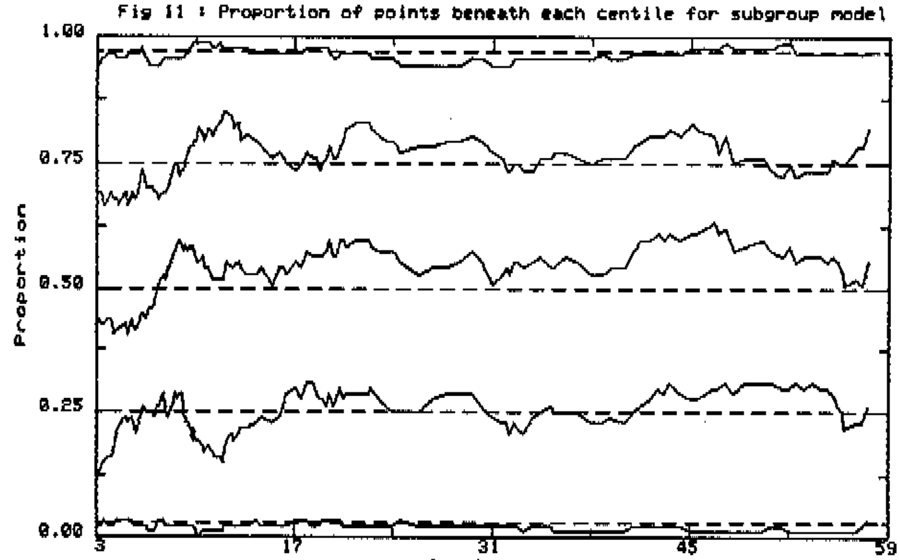
You can use the CHOO, OMIT or SPLIt commands to put the boys and girls centile curves into separate columns.

Just as we did with the model for a single group, we can use the CONT command to check the goodness of fit for the model with subgroups. An extra column is added to the CONT command containing the subgroup information :

SORT C1 C2 C3 C1 C2 C3 (sort data by age)
CONT C2 C1 0.1 C50-C53 C19 C24-C29 5 C3

When counting the measurement points within a box to assess the proportion below a given centile, the boys model is used if the point belongs to a boy and the girls model if the point belongs to a girl. The checking curves produced by this command give an indication of the goodness of fit for both subgroups. Before you use the command CONT you must sort the data in age order. To graph the checking curves, type:

READ C44-C49
0.03 0.25 0.50 0.75 0.97 3
0.03 0.25 0.50 0.75 0.97 59
NEWG
PLTL C24-C29 1 1 1 1 1
PLTL C44-C49 2 2 2 2 2
DISP



You may have noticed that subgroup codes were output along with the checking curves. This is because you may wish to separate the checking curves for boys and girls so you can examine the goodness of fit for each subgroup individually.

You may wish to experiment with other subgroup models.

9.2.3 Using SPLN to obtain Piecewise Polynomials with Subgroups

The worksheet is from section 9.2.1. The syntax of the SPLN command with subgroups is same as with one group with an extra column of codes. Before using SPLN, the raw centiles should be sorted by age, by typing

```
SORT C9 C4-C8 C20 C9 C4-C8 C20
SET C51
12
SPLN 3 2 1 1 0 C51 C50 C4-C9 C14-C19 C20
```

To plot the centiles for both boys and girls, first sort the smoothed centiles by age within subgroup and then use the PLTG command.

```
SORT C20 C9 C14-C18 C20 C9 C14-C18 2
NEWG
PLTG C20 C14-C18 C9
DISP
```

10 Developmental Milestone Curve Fitting Commands

Data for developmental milestones such as 'walking', 'self feeding' etc. are taken and curves are produced which show the percentage of individuals not achieving the milestone at different ages. The commands MILE and DBOX in section 10.1 are for calculating a raw unsmoothed centile curve for longitudinal data and cross sectional data respectively. The command DPOL in section 10.2 is for smoothing the raw centile obtained by the commands in section 10.1.

10.1 The Commands for Obtaining Raw Centiles

Two alternative commands are available for this purpose. This first command is designed to operate on longitudinal data and the second command is designed for cross sectional data.

The command MILE is for longitudinal data.

Suppose data for a group of individuals are recorded on a number of occasions, and on each occasion it is noted whether or not the subject has achieved the milestone. The data for each subject are scanned and the age at which the subject is first recorded as achieving the milestone is noted. Since the subject could have achieved the milestone at any time between the noted age and the previous age the transition age is recorded as being half way between these two ages. If we had data on 50 subjects each measured on 10 occasions we would have 500 data records. After the data had been scanned we would have 50 transition ages. These transition ages are then sorted into ascending age order.

This procedure can be used for mixed longitudinal data where some cases are recorded just once. We assume that all such cases are selected at random with respect to age. In general the ages for such a case will fall between 2 consecutive ages determined from the longitudinal data. The case is inserted into the age sequence at that age. Where such a record is a 'not achieved' record, it does not enter into the computation of 'proportion achieved' at subsequent ages. Where it is an 'achieved' record it is retained at subsequent ages. Thus, for each age, we obtain an estimate of the proportion who have achieved the transition by that age. For longitudinal data where a transition is not observed, either all records are 'not achieved' or 'achieved'. The last or first record only is used and this is equivalent to treating each record as a cross sectional one.

For each of the 50 transition ages the percentage of subjects not achieving the milestone is calculated. In this example, with 50 values, 98% will have not achieved the milestone at the first transition age, 96% at the second and so on. In the event of two or more individuals having the same transition age the percentage not achieving at that age is adjusted accordingly.

The command to form these milestone centile curves for longitudinal data is

```
MILEstone using C C C output into C C
```

The first column contains case identification codes. The second column contains the developmental data and the third column contains ages of subjects when recordings were made. Note that in the second column 0 signifies that the milestone has not been achieved while 1 signifies that it has. The list of transition ages is output to the last column and the percentage of individuals not achieving the milestone at each age is output to the fourth column.

The command DBOX is for cross sectional data.

With purely cross sectional data a different method is employed. The data must be sorted into age order using the SORT command. A box of fixed size is then moved through the data one point at a time (as with the CBOX command). Within each box position the percentage of individuals not achieving the milestone is calculated. This percentage is plotted against the median age value of the points in the box. The command to perform this operation is

DBOX using C C N output into C C

where the first column contains developmental data, in which 0 signifies that the milestone has not been achieved while 1 signifies that it has. The second column contains ages of recordings. Then comes a number N specifying the size of the box as a fraction of total data set size. The output is written to the last two columns which contain, for each box position, the percentage of individuals not achieving the milestone and the age median.

10.2 The Command for Smoothing Curves

The output produced by both the MILEstone and DBOX commands will be very irregular and will require smoothing. Reasonable smoothing can usually be achieved by fitting a polynomial. The command to do this is

DPOL N y data in C x data in C of degree N predicted values to C

If the first number is set to 0 the y data is untransformed. If it is set to 1 a logit transformation is used. The y and x data are the columns of percentage not achieving the milestone and age truncated to 100 and predicted values less than 0 are truncated to 0. If the command is followed by another column then the coefficients of the fitted polynomial are output to this column. In practice a logit transformation generally gives better results.

11 Examples of Developmental Milestone Curve Fitting

This section will describe how to estimate milestone centiles with the commands in section 10.

11.1 Example for Cross Sectional data

This section describes estimation of milestone centiles with the commands DBOX and DPOL for cross sectional data.

In the cross sectional dataset 'lift.dat', with 880 children under 6 months, there are two columns. The first column C1 contains the ages in months. The second column C2 contains the developmental variable 'lift head steady', in which 0 signifies that the milestone has not been achieved while 1 signifies that it has. There are no missing values in either column.

The data must be sorted into age order by typing

SORT C1 C2 C1 C2

and then

DBOX C2 C1 0.15 C3 C4

Where 0.15 is the box size selected. The percentages of individuals not achieving the milestone within each box are output into C3 and the median age values of each box are output into C4. You can plot the percentages against the ages estimated and this plot will help you to select a model. Type

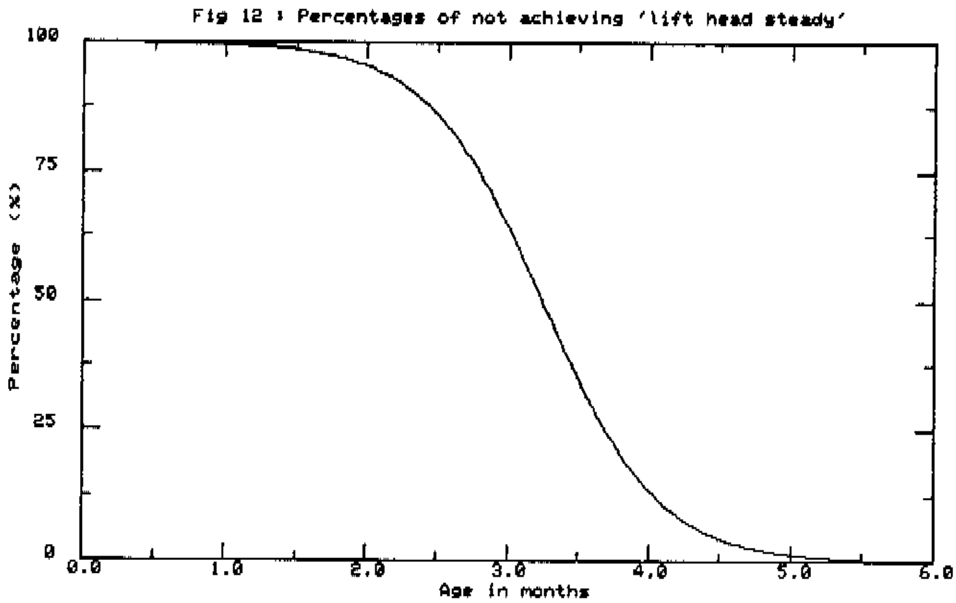
PLTP C3 C4

Suppose you wish to use a logit transformation of the y data and fit a linear model. You can set the first N in DPOL equal to 1 and the second N to 1. Type

DPOL 1 C3 C4 1 C5

To look at the smoothed curve, Figure 12, type

NEWG
PLTL C5 C4
DISP



11.2 Example for Longitudinal data

This section will describe how to estimate milestone centiles with the commands MILE and DPOL for longitudinal data.

Suppose you wish to calculate the milestone centiles for males on the developmental variable 'smile' in the longitudinal dataset 'smile.dat' which contains ID, the developmental variable and age. First you should use the DREAd command to read the data into columns 1 to 3 and deal with missing data with the OMIT command and then type

MILE C1-C5

to produce the unsmoothed centile estimates. You can look at the plot of the estimates by typing

PLTP C4 C5

The plot will help you to select the model, that is the choice of Ns in the DPOL command. For example, the model in which the first N is equal to 1 and the second N is 1 fits a linear model with a logit transformation of the centile proportions in column 5. To do this, type

DPOL 1 C4 C5 1 C6

12 Command Reference

Command	Notes	Section
STOP	**STOP** **Ends a work session.**	3
HELP	**HELP xxxx** xxxx is a command name, display a reminder about a command's syntax.	4.2
DREA	**DREA C...C** **Reads from a file you name (when prompted). Requires no format statement, but columns in data file must be separated by spaces. You must read in all data columns from your data.** **DREAd N C...C** where N is either 1 or 2, the lines of format you will be asked to type. Enables you to skip variables.	4.3
SET	**SET C** Then type numbers into the column. **SET K N** Put a number into a box	4.4
READ	**READ C** **READ C...C** Then you must type the correct number of entries on each line.	4.4
NAME	**NAME** Provides a list of variables, their ranges, and the number of rows and missing values. **NAME C 'VARNAME'** Allows you to name columns.	4.5
PRIN	**PRIN C...C** Shows the contents of the named columns.	4.5
EDIT	**EDIT element M in C to read N** Enables you to alter a particular element in a column.	4.5
CHAN	**CHAN N in C into M and put results into C** **or** **CHAN all values between N and N in C to N results to C**	4.5
CHOO	**CHOO cases with values of N in C, carrying C...C results to C...C** **CHOO cases between N and N in C, carrying C...C results to C and C...C** Selects cases based on one or more specified values.	5.1.1
OMIT	**OMIT cases with values of n in C, carrying C...C results to C...C** **OMIT cases between N and N in C, carrying C...C results to C and C...C** The complement of the CHOO command.	5.1.1
SORT	**SORT values in C putting results into C** **SORT using C, carrying C...C output into C and C...C** **SORT using C C ,carrying C...C output into C C and C...C 2** The first form writes the first column into ascending order. The second form sorts the carrying C...C along with the ascending order of the sort variable (the first column). The third form sorts C...C on two variable.	5.1.2
SPLI	**SPLI C using codes in C results to C...C** Separates a column into subcolumns using codes in an identified column.	5.1.3
GROU	**values in C lower limits in C results to C** Changes a continuous variable into an ordinal categorical variable.	5.1.3
JOIN	**JOIN values in C and C to form C**	5.1.4

ERAS	**ERAS C...C** Erase the named columns.	5.1.5
TIDY	**TIDY** Recovers empty spaces on a worksheet.	5.1.5
WIPE	**WIPE** Removes all data from the worksheet.	5.1.5
SAVE	**SAVE** Prompts for a name of the file into which the current contents of the worksheet will be saved.	5.2.1
RETR	**RETR** Retrieves a saved worksheet	5.2.1
LOGO	**LOGO** Turns recording of your worksession on (and off).	5.2.2
DWRI	**DWRI C...C** Write the specified columns to a file. The program prompts you for a filename.	5.2.3
ENDO	**ENDO** This must appear at the end of a command file.	5.2.4
OBEY	**OBEY** Calls a macro (file) into operation.	5.2.4
HIST	**HIST of C [using interval width N] [midpoints to C, counts to C]** Produces a stem-and-leaf plot of the values in the named column.	6.1
BOXP	**BOXP of data in C...C** Produces a boxplot of the values in the named column(s).	6.2
PLOT	**PLOT y values in C against x values in C** **PLOT values in C between N and N vs values C between N and N** Produces a basic scatterplot.	6.3
MPLO	**MPLO y values in C...C against x values in C** Produces a scatterplot with multiple y variables.	6.3
LPLO	**LPLO values in C against values in C using codes in C** Produces a scatterplot in which different letters are used for the points from different groups.	6.3
PLTP	**PLTP y values in C against x values in C** **PLTP y values in C...C against x values in C [with symbols N...N]** Produces a high-resolution plot of points.	6.4
PLTL	**PLTL using y values in C and x values in C** **PLTL using y values in C...C against x values in C [lines types in N...N]** Produces a high-resolution plot of lines.	6.4
PLTG	**PLTG codes of subgroups in C y values in C...C x values in C** Produces a high-resolution plot of lines for different groups.	6.4
DISP	**DISP** Make high-resolution plots appear on the screen.	6.4
NEWG	**NEWG** Erase contents of a graph being constructed.	6.4
PLTB	**PLTB using bar lengths in C and x values in C** **PLTB for ys in C...C xs in C [shading styles N...N]** Produces a high-resolution bar chart.	6.5
GSAV	**GSAV** Save a graph in a file.	6.7
GRET	**GRET** Retrieves a saved graph.	6.7

LOGT	**LOGT C C** Log to the base 10	7.1
LOGE	**LOGE C C** Log to the base e	7.1
LOGI	**LOGI C C** Ln (p / (1-p)) for $0<p<1$	7.1
ROUN	**ROUN C C** Round to the nearest integer.	7.1
NED	**NED C C** Normal equivalent deviate of q, for $0<q<1$	7.1
SIN	**SIN C C** Sine of an angle in radians.	7.1
COS	**COS C C** Cosine of an angle in radians.	7.1
ADD	**ADD C C** Add the first to the second. More forms can be used.	7.1
SUBT	**SUBT C C** Subtracts the first from the second. More forms available.	7.1
MULT	**MULT N C** Multiply the first by the second. More forms available.	7.1
DIVI	**DIVI N N** Divide the first by the second. More forms available.	7.1
AVER	**AVER values in C** **AVER values in C [results to K K K K]** **AVER N columns C...C [means to C and s devs to C]** Calculates mean(s) and standard deviation(s) etc.	7.2
MOME	**MOME of data in C [store in K K K K]** Provides the first four moments of the data in the column.	7.2
TABU	**TABU mode N factor levels in C [and C]** **TABU mode N factor levels in C and C** **TABU means and sds of values in C factors in C [and C]** The first or the second form prints a means table and the Third form prints a one- or two-way frequency table. For more forms see the section.	7.3
CORR	**CORR values in C and in C [results to K K K K K]** Provides the correlation coefficient and its significance probability. The two means, the standard deviations and the correlation may be stored.	7.4
CORM	**CORM of N variates in C...C and store it in C**	7.4
NPRO	**NPRO for value N** Gives upper-tail probabilities for the value N in relation to the Gaussian distribution.	7.5
TPRO	**TPRO for value N with N df** Gives upper-tail probabilities for the value N in relation to the t distribution.	7.5
CPRO	**CPRO for value N with N df** Gives upper-tail probabilities for the value N in relation to the χ^2 distribution.	7.5
FPRO	**FPRO for value N with N and N df** Gives upper-tail probabilities for the value N in relation to the F distribution.	7.5
REGR	**REGR responses in C on N predictors in C...C [yhats to C coeffs to C]** Performs an OLS regression.	7.6

CBOX **CBOX using C C P C output C...C C [using N] [using C output C]** 8.1

measurement data in C and corresponding age values C
box size as proportion of dataset P
required centiles in C
raw centiles and their age values output to C..C C
[N points between consecutive box positions]
[input group codes in C output group codes to C]

CBOX estimates raw centiles.

N.B. The measurement data must be sorted on age or age within subgroup if centiles for subgroups are being estimated.

CPOL **CPOL using N...N C C...C output to C...C C** 8.2.1

orders of polynomials p, q_0, q_1, q_2,...in N...N
required centiles in C
raw centiles and their ages in C...C
smoothed centiles output to C...C
estimate coefficients output to C

CPOL estimates centile curves with ordinary polynomials.

PCWS **PCWS using C C...C C C...C C output C...C C using [N] [C] [C] [C]** 8.2.2

join points in C
model for each subrange in C...C
required centiles in C
raw centiles and their age values in C...C C
smoothed centiles output to C...C
coefficients for all subranges output to C
[mode N]
[subgroup codes in C]
[centile weights in C]
[age weights in C]

mode	subgroups	centile weights	age weights
1	Yes	No	No
2	Yes	Yes	No
3	Yes	No	Yes
4	Yes	Yes	Yes
5	No	Yes	No
6	No	No	Yes
7	No	Yes	Yes

If you are using centile weights the column containing them must contain one entry for each centile. If you are using age weights the column containing them must contain one entry for each age point.

PCWS produces smoothed centile curves with piecewise polynomials.

N.B. The raw centile data set must be sorted on age, but not on age within subgroup.

SPLN **SPLN using N...N C C C..C C output C..C C [C]** 8.2.3

model of p q_0, q_1,..., in N...N
joint point/points in C
required centiles in C
raw centiles and their age values in C...C C
smoothed centiles output to C...C
coefficients output to C
[subgroup codes in C]

SPLN produces centile curves with piecewise polynomials (spline or grafted polynomials)

N.B. The raw centile data set must be sorted by age, but not by age within subgroup.

CONT **CONT using C C P C C C...C C output C...C C [using N] [using C]** 8.3

measurement data in C and their corresponding age values C
box size as proportion of dataset P
required centiles in C
joint points in C
model for each subrange in C...C
coefficients for all subranges in C
checking curves and their age values output to C...C C
[N points between consecutive box positions]
[subgroup codes in C]

CONT counts percentages of estimated centile points with the coefficients obtained by CPOL or PCWS (but not by SPLN) to check the goodness of fitting.

N.B. The measurement data must be sorted on age, but not on age within subgroup.

CVAL **CVAL using C C C C...C C output C [using C]** 8.3

measurements in C and their corresponding age values in C
joint points in C
model for each subrange in C...C
coefficients for all subranges in C
percentile transformations output to C
[subgroup codes in C]

The CVAL command uses a set of estimated centiles to calculate a percentile value for a given set of measurement points.

You can transform the percentile values to Z scores using the NED command.

N.B. The measurement data must be sorted on age, but not on age within subgroup.

MILE **MILE using C C C output into C C** 10.1

ID of subjects in C
developmental data in C
age in C
percentage of individuals not achieving the milestone to C
related transition ages in C

MILE calculates raw centiles of developmental milestones for longitudinal data.

DBOX **DBOX using C C N output C C** 10.1

developmental data in C
age in C
box size as a fraction of total dataset size in N
percentage of individuals not achieving the milestone to C
age median to C

DBOX calculates raw centiles of developmental milestones for cross-sectional data.

DPOL **DPOL using N, y-data in C x-data in C of degree N predicated values output to C** 10.2

when N=0, y-data is untransformed; when N=1, y-data is logit transformed.
percentage not achieving the milestone in C
age related to y-data in C
degree of polynomial in N
predicated values output to C

DPOL calculates smoothed curve for developmental milestone raw centiles.

13 Command Index

14 References

Goldstein,H., & Pan,H. (1990) Percentile smoothing using piecewise polynomials with covariates. Manuscript submitted for publication.

Healy, M.J.R. (1989). Nanostat users' guide. London: Alphabridge Ltd.

Healy, M.J.R., Rasbash, J., & Yang, M. (1988). Distribution-free estimation of age-related centiles. Annals of Human Biology, 15, 17-22.

Kendall, M., & Stuart, A. (1958). The advanced theory of statistics. (Vol. 1). London: Griffin.

Pan, H., Goldstein, H & Yang, Q., (1990). Non-parametric estimation of age-related centiles over long age ranges. (In press, Annals of Human Biology).

Prosser, R., Rasbash, J., & Goldstein, H (1990). ML3 software for 3-level analysis: Users' guide. London: Institute of Education.

Rasbash, J. (1987). The GRANDSTAT manual. London: Institute of Education/ Geneva: World Health Organization.

Rasbash, J., & Healy, M.J.R. (1988). Data management for growth studies. Annals of Human Biology, 15, 269-273.

= = =