Data Processing for Longitudinal Studies

By H. GOLDSTEIN

Institute of Child Health, London

SUMMARY

An account is given of some data-processing problems which arise in longitudinal studies. Computer programs which have been developed to deal efficiently with such data are described.

1. INTRODUCTION

ALTHOUGH computer programs have been written for the analysis of many different types of survey (see Rowe, 1969), there seem to be none dealing efficiently with data arising from longitudinal studies, that is studies in which repeated measurements are made on the same individuals. Such studies have peculiar problems which arise from the need to relate together the measurements made at several different times, on a single individual. The set of programs described below was written to handle these problems simply and efficiently.

The Department of Growth and Development at the Institute of Child Health has, under the direction of Professor Tanner, been engaged for many years in intensive studies of the physical growth of children. Samples of children are generally measured at yearly intervals on or near their birthdays, at 6-monthly intervals or, during the period of puberty, at 3-monthly intervals. At each examination some 20 body measurements are made, with photogrammetric photographs, and previously, wrist X-rays. Also clinical examinations are made and ratings assigned for the stages of puberty. In addition, the department undertakes various special child-growth studies and laboratory animal-growth experiments from time to time.

The programs have been written by the author together with Margaret Manning and Barry Carter. They have evolved in response to the practical research requirements arising from the work of the department and to the need of researchers for easy use and clear presentation.

The next three sections contain discussions of some of the special problems associated with the data processing of longitudinal studies and are followed by sections describing the computer programs.

2. LONGITUDINAL DATA STRUCTURES

Various attempts have been made to classify longitudinal studies by both the type of sampling scheme adopted and the ultimate aims of the study (Goldstein, 1968). Unfortunately, these classifications are not very relevant to decisions which must be made about the data processing of such studies. Aside from technical considerations about the form of input medium (now almost universally punched cards) or storage medium (usually magnetic tape), the structure of a longitudinal data-processing system will be primarily determined by the number and phasing of examinations on each individual and the number and nature of the measurements made at each examination.

The number and phasing of examinations will determine the structure of data files and the programs to update them. If, for example, all individuals are measured

on the same few occasions, as in most large-scale "cohort" studies, the file structure and associated manipulations are much simpler than if individuals are measured on many occasions, not necessarily the same for each individual. The first case can be handled without much difficulty by simple adaptations of most existing survey programs, whereas the second requires specially designed manipulation techniques.

As for the nature of the measurements, much depends on whether this changes with time. In educational studies, for example, it is inappropriate to ask exactly the same questions or apply exactly the same tests on each occasion, although some common measurements will certainly be made. However, in physical growth studies the same measurements are usually made on each occasion. Although this paper is chiefly concerned with studies of the latter type, many of the remarks and suggestions will also apply to the former.

3. Data Cleaning

The most carefully collected data will never be completely free from gross errors. Such errors may arise at the measuring stage by, for example, misreading an instrument, in transcribing data onto forms, or in punching data onto cards or tape.

Where a variate has a well-defined range (e.g. a score 0–10) a check for illegitimate values is easy. Similarly where the value of one such variate is logically dependent on the value of another, cross-checks can be carried out. With a continuous scale, as for body measurements, checks must identify outlying observations or suspicious values worthy of further inspection. A normally distributed variable, for example, will give 3 per cent of observations outside 2·17 standard deviations either side of the mean. A knowledge of the mean and S.D. therefore will enable us to set up limits beyond which 3 per cent of a sample would be expected to fall if no errors were present, and which will, hopefully, in practice contain a high proportion of actual errors. Clearly, however, not all errors appear as extreme "outliers", and a balance must be maintained between setting limits to catch as many errors as possible and minimizing the number of values to be checked. The narrower the limits, the more suspected errors outside them and the more values to be checked. For single body measurements 3 per cent limits have been found reasonably satisfactory. Generally, the "outlying" errors will be the more important ones to eliminate, since they will have the most marked effect on second- and higher-order moments. In longitudinal studies, we have information on a single individual over time, so that we can have a very much more powerful procedure than is available with one individual measurement at one time, regarded as a random observation from a population. Known functions of variables may be treated in a similar manner. The degree of complexity which can be adopted is limited in practice, and for the detailed physical growth studies at the Institute of Child Health good results have been given by checks using the relationship between pairs of variables "cross sectionally" (i.e. treating an individual as a random member of a population at each age) and using single variables longitudinally (in terms of the rate of growth from one examination to the next). A more comprehensive discussion of error checking is given by Freund and Hartley (1967).

4. Variability of Examination Time

In most longitudinal studies it is usual to try to examine all individuals at predetermined times, for example on their birthdays. Such a procedure, among other things, greatly simplifies the subsequent analysis of the data. If the variability

about such a "target" date is small compared with the changes taking place in the variables being measured, the analysis may still be carried out assuming that the times are exact. The International Children's Centre has adopted acceptable limits for yearly measurements, of two weeks either side of the target date. Very often, however, individuals arrive well outside any such reasonable limits. If we then wish to estimate say, the mean calf circumference of boys aged 8·0 years we may well be faced with the problem that perhaps 20 per cent of our sample is actually measured between 8 years 1 month and 8 years 3 months.

How then can we use these data to estimate the mean at 8 years? If we know the relationships between calf circumference at 8 years and calf circumference at the ages when the boys were actually measured, we could then use the predictions given by these relationships (suitably weighted) to give a more precise estimate of the mean at 8 years. For groups of individuals measured at a fixed number of occasions, such an estimation procedure is given by Patterson (1950). In general, however, we do not know this relationship for arbitrary values of x_i, where the ith individual is measured at $T+x_i$, T being the "target" time. Nor do we have individuals present at time T and $T+x_i$ for all x_i to enable us to estimate these relationships. However, the measurements on all the population in a region around the target date can be used to derive a suitable adjustment. For example, if growth in this region is assumed to be exactly linear for each individual with respect to a known time metameter, this leads to a linear adjustment equation. The parameters of this equation can be estimated from a sample of individuals having measurements at two different times within the region.

An alternative approach is to consider the observed measurements available for each individual near the target date, to fit a curve to these and interpolate at that date. For example, measurements may be made at three time points in an interval around T. A second-order polynomial can be fitted to pass exactly through these points and the value at time T read off. This method will be referred to as the method of individual adjustment.

The first method can be used to adjust individual values only when the unknown parameters in the adjustment equation are available. If the parameters have to be estimated from the sample this method cannot be used until some analysis has been done, and therefore not while the data file is being created. There may also be difficulty in finding a time metameter with respect to which growth is, say, linear.

The second method can be used to adjust individual values at the time the data file is created. Unlike the first method it does not assume a known form of growth curve for each individual. It thus avoids introducing biased estimates for individuals with growth curves which are different from the one assumed. This second method is available in the programs described below.

5. The Computer Programs

A family of programs has been written, each one dealing with a different stage of the processing. They form a compatible set through the use of a common magnetic tape file structure and common control cards. They are as follows:

(1) A general data-editing program, which reads in data, usually from cards, occasion by occasion for each individual. It has facilities for transforming, checking and deriving new variates, and also sets up the initial tape file.

(2) A correction program, which alters or deletes observations on a tape file.

(3) An updating program, which adds new occasions and new individuals to the existing tape file.
(4) A merging program, which merges several tape files onto a single file.
(5) A collation program, which matches and combines occasions for the same individuals located on different tape files.
(6) An output program, which produces various forms of summary statistics, data listings, etc.

As mentioned above, these programs have been written to satisfy the demand of a university research department primarily interested in a few intensive longitudinal growth studies, and the facilities available reflect this. For example, most of the variates are continuous so that summary information is in the form of means, standard deviations, centiles, etc.; in consequence procedures for handling discrete data are admittedly inadequate. Since most users of the programs up to now have had little data-processing experience, the method of specifying summary statistics, etc. has been kept simple, but flexible, with instructions written in English. The creation of the tape files, involving definitions of variates, etc. is fairly complicated and is usually undertaken by an experienced member of the department's statistics and computing section. It is at this stage that derived variates are created.

The following descriptions show the facilities available and are not detailed specifications. Such specifications and copies of program source decks may be obtained from the Statistical Laboratory, Institute of Child Health, 30 Guilford Street, London, W.C.1. All programs are written in Fortran IV for an IBM 7094/1401 or a CDC 6600 computer system. An attempt has been made to keep to ASA standard Fortran with only a few minor differences imposed by the installations used.

Only the general data-editing program and the output program will be described in detail. The functions of the other programs are fairly clear and introduce no new relevant points.

5.1. *General Edit Program*

The edit program creates the initial tape file, which consists of book-keeping information followed by the set of individual data records. The book-keeping information contains a file label, a list of 12-character names assigned to the variates and a list of all control cards specifying the transformations, checks and derivations of new variates.

An individual record consists of a set of occasions, each of which contains a vector of variate values. Within a single file, each occasion contains the same number of variates but the number of occasions per individual may change. Each record is written on tape as a single logical Fortran binary record, the first two words of which contain a serial number and the number of occasions.

For efficiency certain restrictions on the tape format are necessary. Updating and collating demand a well-defined ordering on the file and this is most conveniently done by defining a serial number for each record and requiring these to be in ascending numerical order. This implies that all input cards have a fixed field for the serial number. In addition to the standard input of numeric data on cards, the program will also read data from binary tapes. In particular, these tapes may be created using another special program that accepts data punched on cards in completely free format allowing any hole combination. Card fields are defined using Fortran F or E format, blank fields being recognized as no response and a special value assigned. All the

occasions for one individual constitute one record and are read in consecutively, a new individual being indicated by a change in serial number.

It might be asked why the cards could not be read onto tape in any order and the tape subsequently sorted? This possibility has been considered and rejected on two counts. First, because the addition of a further computing step would involve longer turn-rounds and greater cost, and, secondly, because when corrections have to be made both to the tape and the cards, the cards can most easily be identified if sorted into a reasonable order, especially where this is the same order as on tape. Cards will usually be sorted, within a record, on a key variate such as age, and a simple check for this order can be specified.

As mentioned above, derived variates may be defined using combinations and transformations of the existing variates. The usual arithmetic and logical functions are available together with checking facilities described previously. The program prints error messages if checks fail but takes no correcting action. These features are common to most survey programs. One special feature of this program, however, is its ability to manipulate data across occasions. In the simplest case, we require the calculation of the change (increment) in a measurement from one occasion to the next. We may require the ratio of such increments, for example the ratio of stature increment to age increment to estimate rate of change of stature. Single control cards may be used to specify these and other functions of increments across a specified number of occasions either before or after the occasion being examined. These transformations define new derived variates on this occasion. The setting of conditions for control cards to be obeyed allows a subset of the data only (for instance, boys) to be transformed.

When calculating rates of change or increments, we usually want to determine these for constant time differences. For example, if children are measured on their birthdays and also, sometimes, at 3-monthly and 6-monthly intervals, intervals between occasions will vary from 3 months to 1 year (or more if some examinations are missed). Thus simple rates of change may be estimated over 3-monthly, 6-monthly, 9-monthly or yearly intervals. Their distributions, in particular the variance, will depend on the interval and we must therefore be able to specify which interval we are considering. The control cards for such "longitudinal" transformations allow limits to be set for the increment of a specified variate which inhibit the calculation if exceeded. For example, we can specify that the increment in age should be between 0·9 and 1·1 years, in which case, for each occasion, a search will be made of all previous occasions until one (and only one) is found satisfying this condition. A rate of change using this condition will then be calculated over a 1-year interval.

Another special feature adjusts values to specified times, as described in the section on the variability of examination times. The appropriate control card defines the values of a specified variate (say age) to which a variate (say stature) is to be adjusted (to the nearest integral age, for example). The control card may also contain conditions which define the greatest interval between the adjusted value (of age) and the nearest observed value, the degree of the polynomial to be fitted (currently up to the third degree), the greatest range of observed values over which such a curve is to be fitted or the type of adjustment for the first and last occasions, as well as conditions to be satisfied by other variates on the occasion being considered. As before, these adjusted values become derived variates.

At present the program will accept a maximum record size of 50 occasions per record and 150 variates (including derived variates) per occasion.

5.2. *Correction Program*

The correction program is used to give new values to a specified observation or to delete an observation or a complete record. The corrected file is written onto a new tape.

5.3. *Updating Program*

This program is used to add fresh data to an existing file. These data may be either further occasions for individuals already in the file or new individuals entering a study. The input data on cards are prepared in the same way as for the edit program, with the same format as the cards which were initially used to create the tape file. Occasions and new individuals are inserted in the appropriate position and a new updated tape file created. The control cards which were stored at the beginning of the initial tape file are used to process the new data. It is also possible, using the longitudinal transformation procedures, to modify variates for occasions in the original file. For example, the last occasion in the original record can now have values adjusted to the nearest whole year of age using new later occasions.

5.4. *Merging Program*

This program transfers selected variates from several tape files onto a single tape file. Each file will contain different individuals, and serial numbers may be altered to ensure a unique identification in the merged file.

Often a special analysis requires information to be pooled from several independent studies. The initial creation of a single file greatly cuts down subsequent processing time.

5.5. *Collation Program*

This program collates information from two files containing the same individuals. Collated records form a new tape file. Matching of occasions within a record is specified by setting matching tolerance limits for a chosen variate. Various operating options are allowed. Either only matched occasions may form part of the new file, or all occasions. Multiple matching of a single occasion from one file with many occasions from the other is allowed, this being useful, for example, when a single set of measurements made on a child's parents is to be added to all occasions. Occasions may also be sorted on the value of a chosen variate, and the full range of transformations may be specified by new control cards as for the edit program.

5.6. *Output Program*

This program produces either summary statistics, a file of selected variates written onto tape, a listing of selected data on a line printer or output on cards for input to further analysis programs.

Control cards contain instructions such as *MEAN* or *LIST* and the name of the variate concerned. Conditions may be imposed for the execution of these instructions, and these may be nested as, for example, when a set of instructions is to be carried out for a set of ages for both sexes. Each set may contain up to 18 instructions and these will now be described in detail.

The *LIST* instructions will cause all the variates in the set containing this instruction to be printed. The field width and the number of decimal places may be

simply specified and names assigned to individual variate values. Unless conditions are imposed all occasions will be listed and if required each individual's listing can be made to begin on a new page with specified heading information.

The *MEAN* instruction will cause the calculation of the mean, variance, standard deviation, standard error, range, skewness and kurtosis, for the occasions satisfying the specified conditions.

CORRN instructions within a set cause the correlation matrix and the means for the specified variates to be printed.

SSP instructions within a set cause the sums-of-squares-and-products matrix, correlation matrix and means to be printed and the sum-of-squares-and-products matrix and means to be punched on cards for input to further analysis programs.

The *PCNTL* instruction causes the listing of certain percentiles estimated directly from the sample using linear interpolation, together with means and standard deviations. For clear presentation, if a condition is repeated for more than one value of a variate (usually age) the listing is organized so that all the conditions for a single variate come together on a page. Cards are also punched out containing the value of the condition variate and the 3rd, 10th, 25th, 50th, 75th, 90th, 97th percentiles so that these can be input to a program which will plot the percentiles against the value of the condition variate (age).

The *WRITE* instruction causes all the variates thus specified within a set to be written onto a scratch tape. Several programs have been written for special analyses using this tape. One carries out the procedure mentioned above (Patterson, 1950) for producing efficient estimates of means from mixed longitudinal data. This tape may also serve as input to a general multiple-regression program which includes graph-plotting facilities.

In some respects the output program is unsatisfactory. It has no facilities for producing contingency tables or for making transformations before instructions are executed. It is also somewhat tedious to have to transfer information via tape or cards between this program and other programs, particularly for often-used programs, such as the graph-plotting program and the program for producing efficient estimates of means. Unfortunately, with the computer system at present in use, the existing program cannot be made to yield sufficient storage to accommodate any additions, at least not without sacrifices in flexibility and speed. However, the transfer to a CDC 6600 now taking place will solve these problems.

Acknowledgements

I would like to express my gratitude to Professor J. M. Tanner for providing the original stimulus and continuing support for the development of these programs. This work was made possible by a grant from the Nuffield Foundation to the Department of Growth and Development at the Institute of Child Health.

References

Freund, R. J. and Hartley, H. O. (1967). A procedure for automatic data editing. *J. Amer. Statist. Ass.*, **62**, 341–352.

Goldstein, H. (1968). Longitudinal studies and the measurement of change. *The Statistician*, **18**, 93–117.

Patterson, H. D. (1950). Sampling on successive occasions with partial replacement of units. *J. R. Statist. Soc.* B, **12**, 241–255.

Rowe, B. C. (1969). *Survey Analysis on Computers.* Abstract of B.C.S. Symposium, Datafair, 1969. London: British Computer Society.