Pergamon

Comput. Biol. Med. Vol. 24, No. 4, pp. 323–328, 1994

0010-4825/94 $7.00 + 0.00

0010-4825(94)00019-0

A PC PROGRAM FOR CLASSIFICATION INTO ONE OF SEVERAL GROUPS ON THE BASIS OF LONGITUDINAL DATA

EMET D. SCHNEIDERMAN,* STEPHEN M. WILLIS* and CHARLES J. KOWALSKI†

* Department of Oral and Maxillofacial Surgery and Pharmacology, Baylor College of Dentistry, P.O. Box 660677, Dallas, TX 75266–0677, U.S.A.; and † Department of Biologic and Materials Sciences and The Center for Statistical Consultation and Research, University of Michigan, Ann Arbor, MI 48109, U.S.A.

(*Received* 4 *January* 1994; *in revised form* 7 *July* 1994; *received for publication* 21 *July* 1994)

Abstract—A stand-alone, menu-driven PC program, ZCLASS, written in GAUSS386i, for classifying subjects into one of several distinct, existing groups on the basis of longitudinal data is described, illustrated, and made available to interested readers. The program accepts data from studies where common times of measurement are planned, but missing data are accommodated in that one or more measurement sequences may be incomplete.

Longitudinal data Polynomial growth curves Discrimination Classification

INTRODUCTION

In a recent paper [1] we described, illustrated, and made available a PC program, written in GAUSS386i, for diagnosing abnormal growth, growth velocity and acceleration given longitudinal data from a normative sample of size N and from an $(N+1)$st individual being evaluated. The purpose of the present paper is to extend this methodology to the case where we have longitudinal observations from G groups and wish to classify a new subject into one of the groups on the basis of his/her observed pattern of growth. The method should prove useful in assessing the growth pattern of an individual subject or patient; if several growth patterns are known for a population, perhaps some of which have undesirable outcomes, it would be of prognostic value to be able to classify an individual subject into one of the groups with a known level of confidence. Information of this sort could be used in the pre-emptive treatment of certain growth disorders. It also has potential applications in classifying patients with regard to how they respond to different therapies over the course of time. The technique is due to Zerbe [2]. It is assumed that the study is planned so that individuals will be measured at the same points in time, but missing data are allowed. The variable whose growth is being monitored need not have a Gaussian (normal) distribution.

In the next section we summarize Zerbe's procedure. We then illustrate the technique and the use of our program on a data set consisting of longitudinal measurements of stature in $G=3$ groups of individuals. Information on obtaining copies of the program is provided in the Appendix.

ZERBE'S PROCEDURE

Consider G groups of subjects, n_g in the gth group, $\Sigma n_g = N$, and an $(N+1)$st individual to be classified into one of the groups. Following the procedure detailed in refs [1–4], we obtain the average distances

$$\bar{d}_{N+1}(1), \bar{d}_{N+1}(2), \ldots, \bar{d}_{N+1}(G) \quad (1)$$

of the $(N+1)$st individual from the individuals in each of the G groups. These are measures of the $(N+1)$st individual's similarity to the members of the G groups under consideration; the smaller the value of $\bar{d}_{N+1}(g)$, the more similar he/she is to the members of group g. We also compute the proportions of individuals having values $\bar{d}_i \leq \bar{d}_{N+1}(g)$ for $g = 1, 2, \ldots, G$ and $i = 1, 2, \ldots, n_g$, i.e. the proportions of individuals in each of the G groups which are closer together than the $(N+1)$st individual is to members of that group. We denote these proportions by

$$\gamma_1, \gamma_2, \ldots, \gamma_G \tag{2}$$

and note that the smaller the value of γ_g, the more centrally located is the growth curve for the $(N+1)$st individual relative to the growth curves of the members of the gth group. Indeed, a value of $\gamma_g = 0$ would imply that the $(N+1)$st individual was the most central (or most typical) member of that group. It is natural, then, to classify the $(N+1)$st individual into the group with the smallest value of γ_g. Note that while both (1) and (2) are measures of similarity, the classification is made on the basis of (2). A given $\bar{d}_{N+1}(g)$ can be large even if $\gamma_g = 0$; and $\bar{d}_{N+1}(g)$ can be small even if $\gamma_g = 1$. One need only imagine a number of growth curves for group 1 which are relatively variable but within which the curve for the $(N+1)$st individual is centrally located; the curves for group 2, however, are bunched tightly together, and close to that for the $(N+1)$st individual, but the curve for the $(N+1)$st individual lies just outside the scatter. The process described above would lead to classification into group 1.

Our program computes the average distances (1) for the polynomial growth curves, the growth velocity curves, and the growth acceleration curves [4]. We output (1) and (2) so that investigators can order the groups with respect to their similarity to the $(N+1)$st individual and get some feeling for the confidence with which this individual was classifed. We also plot the average growth curves (AGCs) for each of the groups and the growth curve for the $(N+1)$st individual.

AN EXAMPLE

In order to illustrate our program we consider three samples of children living in Guatemala, which were studied in depth by Bogin *et al.* [5]. The children comprising these samples differ in socioeconomic status (SES) and ethnicity: one is of high SES Ladino children (G_1); the second is of low SES Ladino children (G_2); and the third is of low SES Mayan children (G_3). There are 20 children in each group and we consider their growth in stature, this being measured $T = 6$ times at ages 7, 8, 9, 10, 11 and 12 years. This same data set was used in ref. [6] to study the tracking behavior of these individuals and in ref. [7] to illustrate the use of the Potthoff–Roy analysis to contrast the AGCs in the groups. We suppose a new Guatemalan child presents with measurements 112, 117, 125, 130, 137 and 145 (cm). For purpose of comparison, the mean values in the three groups at each time of measurement are shown below [7]:

G_1: 122.300 127.800 134.100 139.925 146.075 153.000

G_2: 113.675 118.675 123.800 129.025 134.575 140.375

G_3: 111.450 116.175 121.450 126.600 131.775 136.800.

The program expects an ASCII data set with 61 rows and (at least) seven columns. One of the columns in the data set will be the group indicator variable; six columns are reserved for the values of the repeated measurements. These variables can be anywhere in the data set, but the repeated measurements must occupy consecutive columns. The user is prompted for the column number containing the group indicator variable; and for the column numbers of the first and last response variables.

The group variable *must be the* positive integral values $1, 2, \ldots, G+1$. If there are G groups, the subject to be classified should be assigned group number $G+1$. In the context of the above example, if the group variable is in column # 1 followed by the longitudinal observations, the 61st row of the data set would be

4 112 117 125 130 137 145

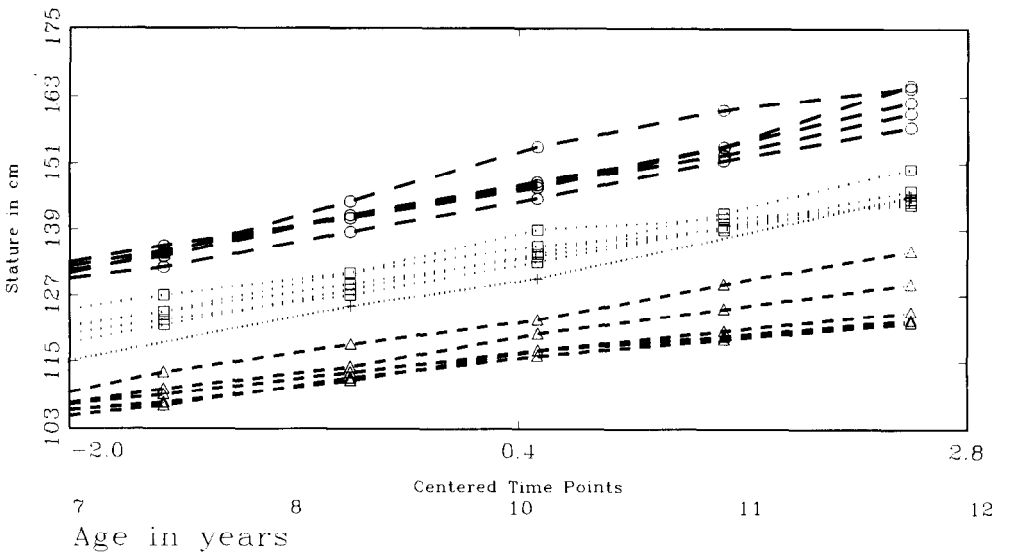

Fig. 1. Plot of the growth profiles of three groups of children (○, □, △) and a "new" child to be classified (+).

and the user would indicate that the group variable is in column 1; the first and last response variables in columns 2 and 7, respectively.

The program is invoked by the command *gsruni zclass* and the user first determines either a common degree, D, to be used to fit each of the growth profiles, or a series of degrees $D_1, D_2, \ldots, D_{N+1}$ specific for each individual using the methods given in refs [1, 3, 4]. In the former case, the distances are computed directly as outlined above; in the latter, the user can choose between augmenting with zeros or (re)fitting everyone to $D_{\max}$. Augmentation involves adding enough zeros so that each case has the same number of regression coefficients, as is required for computing the distances [1–4]. For example, if a given $D=1$ and $D_{\max}=3$, the vector of regression coefficients for the subject in question is made 4×1 by appending two zeros to the vector containing the intercept and the coefficient of the linear term.

In our example, taking $D=2$ for each subject as in [7], we find

$$\bar{d}_{N+1}(1)=22.037,\ \bar{d}_{N+1}(2)=11.834,\ \bar{d}_{N+1}(3)=11.467,$$

from which

$$\gamma_1=76.18,\ \gamma_2=20.00,\ \gamma_3=40.00$$

and it is seen that the growth curve of the new child most resembles the growth curves in G_2. That this is a reasonable conclusion is supported by comparing the values for the new child with those of the AGCs, and by plots produced by our program which show the fitted growth curves for each subject, color-coded for group membership. While the use of color allows us to see this quite clearly on the screen even for $N=61$ cases, the black-and-white plot is relatively cluttered and so is not reproduced here. To illustrate the usefulness of this plot, we show the result for a subset of the original data, five children from each group, in Fig. 1. It is seen that the new child (whose growth profile is indicated by + signs) is intermediate to those of G_2 (□) and G_3 (△), but closer to G_2.

Analogous results are also provided for the velocity and acceleration curves [4]. While not shown here, they do, in this example, support the notion that the new child is most likely from G_2, or at least from one of the low SES groups. This, however, is not *guaranteed* to happen. A given subject can be classified into different groups depending on which curve is in question. For example, a child who is large for the first several years

but decelerating may be classified as from G_1 on the basis of his/her growth curve, but as from another group when accelerations are considered.

DISCUSSION

In the simple example considered above, each individual was measured at the same points in time and there were no missing data. In actuality, the program is more general than this. It can also be used in the cases when common times of observations are planned, but some data are missing. How this works is detailed in refs [1, 3, 4]. For present purposes, it suffices to note that the data set is prepared using periods (.) to represent missing data, for both members of the existing groups and the subject to be classified. For example, if the new child considered above had missing observations at 8 and 11 years of age, we would type the data as

$$4\ 112\ .\ 125\ 130\ .\ 145.$$

To give at least a rough idea of how missing data can affect the results of the analysis, we note that when the above data are used,

$$\bar{d}_{N+1}(1) = 21.528,\ \bar{d}_{N+1}(2) = 11.898,\ \bar{d}_{N+1}(3) = 11.617$$

$$\gamma_1 = 76.19,\ \gamma_2 = 25.00,\ \gamma_3 = 40.00.$$

It is seen that neither the distances nor the percentages are effected to any great extent, and that the conclusions reached are identical.

We should also note that we have used the word classification to describe the process of assigning a new individual to one of G existing groups. Others, e.g. Lachenbruch [8], use the term discrimination and refer to the study of this process as discriminant analysis. Still others, e.g. Kendall [9], use classification synonomously with clustering or cluster analysis, where one uses measurements on each of N individuals in an attempt to identify subgroups with similar characteristics. Our choice of terminology makes no particular point: we have avoided the use of the term discrimination mainly because this refers to a well-defined body of theory focusing on the use of Mahalanobis' distance in the context of multivariate normal distributions to develop rules for optimal allocation (where optimal is variously defined in terms of the prior probabilities associated with each of the groups, and the costs of misclassification). It also generally involves measurements on a number of different attributes, rather than repeated measures of the same characteristic. Zerbe's method, in contrast, is more of an *ad hoc* procedure which may be used to facilitate discrimination given longitudinal data with missing observations for the individuals under consideration. Moreover, the allocation rule suggested by Zerbe, while intuitively satisfying, can lay no claim to optimality with respect to any well-defined criterion. In any case, Lachenbruch [8] defines, "The basic problem of discriminant analysis is to assign an observation, x, of unknown origin to one of two (or more) distinct groups on the basis of the value of the observation", so it is clear that we are here in fact considering the same problem. Implementation of the obverse problem of cluster analysis is considered in ref. [3].

Finally, we note that other approaches to the problem of discrimination on the basis of time-dependent data are possible. In ref. [10], polynomials were fitted to the growth curves in each group, and the resulting regression coefficients were used in a conventional discriminant analysis. Maximum likelihood estimation was employed, and this required an iterative computational procedure. The classical approach was also used by ref. [11], but in the context of exponential decay curves modelled by an autoregressive stochastic process of first order. An approach to classification using repeated sets of multiple measurements over time is described in ref. [12]. Here, two groups were considered (survival and death). At each time point, a conventional discriminant function was calculated. This function then served as a (univariate) measurement over time. Lines were fitted to this measurement and the resulting slope and intercept were used as the basic data in another discriminant function analysis. Work on implementing

procedures such as these continues so that they may be compared to one another and to Zerbe's technique. In the meantime, we offer our program as a potentially useful aid to classification; one that can be used to obtain relevant information when, as so often happens in longitudinal research, missing data must be contended with.

SUMMARY

A method for classifying subjects into distinct, existing groups on the basis of longitudinal observations has been described, illustrated, implemented, and made available. Subjects can be classified using either their growth curves, their growth velocity curves, or their growth acceleration curves. It is assumed that the study giving rise to the data is planned, so that subjects will be measured at a common set of time points, but missing data are allowed. Given an individual to be classified, the key output from the program consists of the proportions of subjects in each of the groups whose average distance from the members of their group is less than or equal to the corresponding average distance for the subject being classified. The smaller this proportion, the more the new individual resembles the members of that group, and the classification rule is to assign the unknown individual to that group for which this proportion is smallest.

Acknowledgement—Supported by DE08730 from the National Institute of Dental Research.

REFERENCES

1. E. D. Schneiderman, S. M. Willis and C. J. Kowalski, A PC program for diagnosing abnormal growth, growth velocity and acceleration from longitudinal observations, *Int. J. biomed. Comput.* **35**, 247–254.
2. G. O. Zerbe, A new nonparametric technique for constructing percentiles and normal ranges for growth curves determined from longitudinal data, *Growth* **43**, 263–272 (1979).
3. E. D. Schneiderman, S. M. Willis and C. J. Kowalski, Clustering on the basis of longitudinal data, *Comput. Biol. Med.* **23**, 399–406 (1993).
4. E. D. Schneiderman, S. M. Willis and C. J. Kowalski, PC program for estimating polynomial growth, velocity and acceleration curves when subjects may have missing data, *Int. J. biomed. Comput.* **33**, 249–265 (1993).
5. B. Bogin, T. Sullivan, R. Hauspie and R. B. MacVean, Longitudinal growth in height, weight, and bone age of Gautemalan Ladino and Indian school children, *Am. J. Hum. Biol.* **1**, 103–113 (1989).
6. E. D. Schneiderman, C. J. Kowalski and T. R. Ten Have, A GAUSS program for computing an index of tracking from longitudinal observations, *Am. J. Hum. Biol.* **2**, 475–490 (1990).
7. T. R. Ten Have, C. J. Kowalski, E. D. Schneiderman and S. M. Willis, A PC program for performing multigroup longitudinal comparisons using the Potthoff-Roy analysis and orthogonal polynomials, *Int. J. biomed. Comput.* **30**, 103–112 (1992).
8. P. A. Lachenbruch, *Discriminant Analysis*. Hafner, New York (1975).
9. M. G. Kendall, *Multivariate Analysis*, 2nd Edn. Griffin, London (1980).
10. S. P. Azen and A. A. Afifi, Asymptotic and small-sample behaviour of estimated Bayes rules for classifying time-dependent observations, *Biometrics* **28**, 989–998 (1972).
11. K. Ulm, Classification on the basis of successive observations, *Biometrics* **40**, 1131–1136 (1984).
12. S. P. Azen and A. A. Afifi, Two models for assessing prognosis on the basis of successive observations, *Math. Biosci.* **14**, 169–176 (1972).

About the Author—EMET D. SCHNEIDERMAN received a B.A. and M.A. in Anthropology from Northwestern University in 1978, and Ph.D. in Biological Anthropology from The University of Michigan in 1985. While at The University of Michigan he was affiliated with the Center for Human Growth and Development and began conducting research in the area of craniofacial growth. In collaboration with Joseph Mudar, Dr Schneiderman developed an integrated software system for the analysis of cephalometric radiographs (X-rays of the head). While on the Orthodontics Faculty of the University of Detroit School of Dentistry from 1985 to 1988, Dr Schneiderman created a computerized cephalometry laboratory. In 1988 he went to the Baylor College of Dentistry in Dallas where he is associate professor and director of research for the Department of Oral and Maxillofacial Surgery and Pharmacology, and he played a major role in initiating a new Ph.D. program in craniofacial biology at Baylor in 1993. Dr Schneiderman and co-investigator Dr Charles Kowalski have been funded by NIH/NIDR from 1988 to 1994 to conduct the biostatistical research from which this paper issued. Dr Schneiderman has more than 70 publications including two chapters and the monograph, *Facial Growth in the Rhesus Monkey*, published by Princeton University Press in 1992.

About the Author—STEPHEN M. WILLIS received the B.S. degree in Mathematics from the University of Texas at Arlington in 1987. Mr Willis has over 15 years' experience in clinical

toxicology and is currently operations manager of a regional toxicology laboratory in Dallas. He is also the lead programmer/systems analyst for the NIH/NIDR grant on longitudinal statistical methods with Drs Kowalski and Schneiderman. Mr Willis has played a major role in the development of user-friendly interfaces for programs that have broad applications in the biomedical sciences, and has coauthored more than 20 scientific publications concerned with these programs. Mr Willis is also an accomplished amateur astronomer.

About the Author—CHARLES J. KOWALSKI received the B.S. in Mathematics from Roosevelt University in Chicago in 1962, M.S. in Statistics from Michigan State University in 1965, and Ph.D. in Biostatistics from The University of Michigan in 1968. He then joined the faculty of the Department of Oral Biology at the University of Michigan School of Dentistry. Dr Kowalski served as the assistant director of the university's Statistical Research Laboratory from 1971 to 1978 and research scientist at the Dental Research Institute from 1978 to the present, and directed the institute's biometrics laboratory. He has been full professor of dentistry and statistician at The University of Michigan since 1978. At various times Dr Kowalski has served as a consultant to the National Football League, Park, Davis and Co., Nijmegen University, Lancaster Cleft Palate Clinic, the Department of Antiquities of the University of Alexandria in Egypt, the U.S. Veterans Administration and the Eastman Dental Center. Dr Kowalski has published more than 200 scientific papers, including numerous chapters and the book *A Mixed-Longitudinal Interdisciplinary Study of Growth and Development*, published by Academic Press in 1979. Dr Kowalski's research has focused on the application of statistical methods to dental and oral research with special emphasis on measurement processes, their validity, reliability and calibration. Longitudinal data analysis and the computer implementation of polynomial growth curve models have also been and continue to be a major thrust of his research. Drs Kowalski and Schneiderman have been funded by NIH/NIDR from 1988 to 1994 to study and implement biostatistical methods for the analysis of longitudinal data in the form of user-friendly microcomputer programs.

APPENDIX

A full set of PC programs for longitudinal data analysis, including this program, can be obtained on 5.25″ or 3.5″ diskettes (please request type) by sending $25 to defray the cost of handling and licensing fees. These programs require an 80386 or 80486 based personal computer (PC) running the MS-DOS operating system (version 5.0 or higher is recommended, although versions as low as 3.3 will suffice). 80386 computers *must* also be equipped with an 80387 math coprocessor. At least 4 Mb of memory is required, and must be available to GAUSS386i, i.e. not in use by memory resident programs such as *Windows*. EGA or VGA graphic capabilities are required to display the color graphics; VGA or SVGA is suggested to display optimally the graphic results. Runtime modules are supplied with the programs so that no additional software (i.e. compiler or interpreter) is required to run these programs. One can create and edit ASCII data sets for use by these programs using the full screen editor supplied with MS-DOS version 5.0. The programs are written and compiled using GAUSS386i, version 3.0, require no additional installation or modification, and are run with a single command. When requesting the programs, address inquiries to the corresponding author and make cheques payable to Baylor College of Dentistry.