Int J Biomed Comput, 33 (1993) 249–265
Elsevier Scientific Publishers Ireland Ltd.

PC PROGRAM FOR ESTIMATING POLYNOMIAL GROWTH, VELOCITY AND ACCELERATION CURVES WHEN SUBJECTS MAY HAVE MISSING DATA

EMET D. SCHNEIDERMAN[a], STEPHEN M. WILLIS[a] and CHARLES J. KOWALSKI[b]

[a]*Department of Oral and Maxillofacial Surgery, Baylor College of Dentistry, 3302 Gaston Ave, Dallas, TX 75246 and* [b]*Department of Biologic and Materials Sciences and the Center for Statistical Consultation and Research, The University of Michigan, Ann Arbor, MI 48109 (USA)*

(Received December 23rd, 1992)
(Accepted February 8th, 1993)

A stand-alone, menu-driven PC program, written in *GAUSS386i*, for estimating polynomial growth, velocity, and acceleration curves from longitudinal data is described, illustrated and made available to interested readers. Missing data are accommodated: we assume that the study is planned so that individuals will have common times of measurement, but allow some of the sequences to be incomplete. The degrees, D_i, adequate to fit the growth profiles of the N individuals are determined and the corresponding polynomial regression coefficients are calculated and can be saved in *ASCII* files which may then be imported into a statistical computing package for further analysis. Examples of the use of the program are provided.

Key words: Polynomial growth curves; Longitudinal studies; Missing data; PC program

Introduction

A number of methods for estimating and comparing the average polynomial growth curves (AGCs) in one or more groups of individuals exist when subjects are measured at identical times, modeled with polynomials of the same degree and when a multivariate normal distribution for the repeated measurements can be assumed. Among the available one-sample methods are those given in Refs. 1–3; the best known of the G-sample procedures is that of Ref. 4. These have been implemented and the procedures are described in Refs. 5–10. Other applications, not concerned solely with AGCs, are considered in Refs. 11–14. These methods and programs are able to provide considerable insight into growth and developmental processes *whenever the conditions mentioned above are satisfied*, but practical circumstances often preclude their application. The assumption of common times of measurement is especially troublesome: individuals invariably miss one or more appointments. Excluding such individuals from the analysis wastes information; estimating the missing values so that they may be included is difficult and may introduce additional (strong) assumptions into the analysis which many researchers would rather avoid.

Correspondence to: Emet D. Schneiderman, Department of Oral and Maxillofacial Surgery, Baylor College of Dentistry, 3302 Gaston Ave, Dallas, TX 75246, USA.

The purpose of the present paper is to describe, illustrate and make available a stand-alone, menu-driven PC program which, given longitudinal measurements on each of N individuals, can be used to estimate and save the polynomial regression coefficients for each individual's growth, velocity and acceleration curves. These coefficients can be saved in *ASCII* files and may subsequently be imported into any statistical computing program which accepts such data sets, e.g. SAS and SYSTAT. The N individuals can comprise $G \geq 1$ groups and missing data are accepted. We assume that the study is planned so that every individual will be measured at the same points in time,but that one or more of the subjects miss one or more of the scheduled appointments. Special attention is paid to the determination of the degree, D_i, adequate to fit the growth profile of the ith individual. The technique is due to Zerbe [15]; we maintain the notation established in Refs. 5–14.

Zerbe's procedure

We consider longitudinal data sets of the form

$$\mathbf{X} = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1T} \\ x_{21} & x_{22} & \cdots & x_{2T} \\ \vdots & \vdots & \cdots & \vdots \\ x_{N1} & x_{N2} & \cdots & x_{NT} \end{bmatrix} \tag{1}$$

where x_{ij} denotes the value of the measurement under consideration for individual i ($i = 1,2,...,N$) at time t_j ($j = 1, 2, ...,T$). In (1), some of the x_{ij} may be missing, but we assume that such missing data points are 'missing at random,' i.e. that occurrences of missing data are not related to the values of neighboring measurements [16]. We assume that a polynomial of some degree (to be determined) adequately fits the growth profile of each of the N individuals.

Fit a polynomial (the procedure we use is described below) of degree D_i to the data from each of the N individuals and let $D = \max\{D_i\}$. Estimate the growth curve for individual i by

$$\hat{x}_i(t) = [1, t, t^2, \ldots, t^D]\,\hat{\tau}_i \tag{2}$$

where $\hat{\tau}_i$ denotes the estimated polynomial regression coefficients

$$\hat{\tau}_i = (\mathbf{W}_i'\mathbf{W}_i)^{-1}\,\mathbf{W}_i'\mathbf{x}_i \tag{3}$$

augmented with $D - D_i$ zero elements. Thus while a particular individual may require but a $D_i = 2$ degree equation to adequately describe his/her data, if $D = 5$, the 3×1 vector computed in Ref. 3 is made to be 6×1 by adding $D - D_i = 3$ zero elements (a quadratic is a quintic with three zero coefficients). An alternative approach [17] is to refit each individual to $D = \max\{D_i\}$. This is considered below. The reason that one or the other of these approaches is necessary is that virtually every statistical computing package requires that the input data for each individual

be of the same dimension and structure, i.e. that each case have the same number of variables ($P = D + 1$) arranged consistently across the columns of the data set. Most of these packages 'allow' missing data, but cases with missing data are simply excluded from any analysis involving the variables in question. In the problem under consideration, all N cases provide information about all P regression coefficients, even though the values of some of these coefficients may be small, or even zero, for some individuals.

In Ref. 3, $\mathbf{W}_i$ is the within-individual (time) design matrix specific to the ith individual. For example, consider a study in which the planned times of measurement are t_1, t_2, t_3, t_4, t_5 and the ith individual was not measured at t_4. Using the successive-powers-of-t form of this matrix [5], if a quadratic equation is to be fit to this individual's growth profile, the appropriate design matrix would be

$$\mathbf{W}_i = \begin{bmatrix} 1 & t_1 & t_1^2 \\ 1 & t_2 & t_2^2 \\ 1 & t_3 & t_3^2 \\ 1 & t_5 & t_5^2 \end{bmatrix} \tag{4}$$

For growth velocity, (2) is replaced by its element-by-element derivative

$$\hat{v}_i(t) = [\,0, 1, 2t, 3t^2, \ldots, Dt^{D-1}\,]\hat{\tau}_i \tag{5}$$

and for acceleration by

$$\hat{a}_i(t) = [0, 0, 2, 6t, \ldots, D(D-1)t^{D-2}\,]\hat{\tau}_i \tag{6}$$

We should note at this point that since we allow missing data, it is necessary to consider the choice of an interval, say $[a, b]$, over which the polynomial regression coefficients will be presumed valid. Since common times of measurement $t_1, t_2, \ldots, t_T$ were planned, it is perhaps natural to choose $a = t_1$ and $b = t_T$, but when subjects present with possibly non-overlapping observation times, it is important to ensure that the regression coefficients for individuals are comparable, i.e. presumed to be valid over the same $[a, b]$. In our program, the user has control over $[a, b]$ so that the choice of a biologically important time period may be balanced with the need for the fitted curve to fairly represent the data over this interval. Plots and goodness-of-fit statistics are provided to facilitate this choice. The plots, as will be demonstrated later, are especially useful in guarding against problems associated with extrapolation. This is considered in greater detail in the next section, where the practical problems in applying Zerbe's procedure are addressed.

Determination of the D_i

The determination of the D_i, the lowest degree polynomial adequate to fit the data for the ith individual, requires some comment. While the theory behind tests for sets of polynomial regression coefficients, e.g. tests of hypotheses of the form

$$\text{H: } \tau_{Q+1} = \tau_{Q+2} = \ldots = \tau_P = 0 \tag{7}$$

is relatively well-known and accessible (Ref. 18, p. 102; Ref. 19, p. 306) implementation is not entirely straightforward and we provide enough detail to indicate why certain decisions concerning program operation were made and to facilitate its effective use. Tests of (7) are based on the fact that

$$F^* = \frac{\text{SSE (R)} - \text{SSE (F)}}{P - Q} \Big/ \frac{\text{SSE (F)}}{T - \text{P}} \sim F(P - Q, T - P) \tag{8}$$

where SSE (R) denotes the sum of squares for error of the 'reduced model' under H and SSE (F) the corresponding sum of squares for the 'full model,' when all P regression coefficients are included. This terminology is consistent with that used in Refs. 19 and 20. In (7) and (8) we have maintained the notation established in Refs. 5–14. P is the number of parameters in the full model which is of degree $D = P - 1$; Q is the number of parameters in the reduced model which is of degree $Q - 1$. One could, in theory, then simply allow the user to specify P and Q thus determining D_i on the basis of the test (8). This is essentially what we do, with certain modifications dictated by the following problems considered in turn: (i) high degree polynomials; (ii) specification of P and Q; (iii) stepping up vs. stepping down; (iv) refitting vs. augmentation with zeros; (v) extrapolation and the choice of $[a, b]$.

High degree polynomials

Zerbe [15], in an example involving $T = 30$ longitudinal weight measurements on $N = 10$ girls from 4 to 18 years of age at half-yearly increments, started by fitting $D = 13$ degree polynomials to each growth profile, apparently guided by the contention in Ref. 21 that 'high degree polynomials are often appropriate'. He then used a 'step up best fit procedure' to see if lower degrees might be used for some of the girls, augmenting with zeros when this was the case, so that the final vector of estimated regression coefficients for each girl was 14×1. In order to allow users to use high degree polynomials, we found it necessary to allow alternatives to the use of the successive powers-of-t form of the **W** matrix as shown in (4) and used in Ref. 5. In order to avoid multicollinearity problems (Ref. 19, p. 295) and computational problems due to round-off error, we allow the user to *center* the data (Ref. 19, p. 315) by using the transformed time scale $t_j^* = t_j - \bar{t}$ for $j = 1,2, \ldots, T$ where $\bar{t}$ is the mean of the t_j. Plots produced by our program show both the centered and original time points. We might have used orthogonal polynomials [20] to circumvent the multicollinearity problem, but simply centering the data represents a reasonable compromise between computational accuracy on the one hand and interpretation of the regression coefficients on the other. It is also true that for certain patterns of missing data we would need to call the program ORPOL [20], which computes the orthogonal coefficients, relatively large numbers of times and this would significantly increase the running time of the current program. Note that when the data are centered, τ_1, the 'intercept,' is the value of the growth curve at $t^* = 0$, i.e. in the middle of the range of observations on the original time scale.

In any event, while our program can accommodate high degree polynomials, the

reader should note that choosing a high degree leaves but relatively few degrees of freedom for the denominator number of degrees of freedom for the F-statistic in (8). The large critical values of F in this case make it difficult to reject (7). This does *not* explain the high final degrees that Zerbe [15] found necessary to fit his data, nor the fact that he found coefficients like $\hat{\tau}_{12} = 6.7081335 \times 10^{-9}$ to be 'significant,' but it does point to the necessity to not rely on automated, 'blind' methods of producing the D_i. Our program does do tests (described below), but also provides alternative guides to the final D_i values to be used in estimating the regression coefficients. The final choice is left to the user.

Specification of P and Q

Zerbe [15] and Dawson et al. [17] both, implicitly at least, considered that specification of a single pair of values (Q, P) was sufficient for all of the N individuals under consideration, i.e. that there was no need for Q and P to be individual-specific. We found that this approach was often precluded by patterns of missing data that could be expected to occur routinely in practice. If $T_1, T_2, \ldots, T_N$ are the numbers of times of measurement for the N individuals, and if a single value of P is to be used for all individuals, one is required to select $P \leq T_{\min} - 1$, where $T_{\min}$ denotes the smallest of the T_i. This would mean that if, e.g., $T_{\min} = 4$ (as was the situation in Ref. 17) that the largest single P that could he used is $P = 3$, i.e. no degree higher than a quadratic could be contemplated; and, as indicated above, tests based on this degree would be compromised by degree of freedom problems. We were thus lead to seek appropriate values of D_i on an *individual basis*. Unless overridden by the user, individuals are considered one-at-a-time, the form of the test for specification being based solely on the data for the subject under consideration. This is somewhat time-consuming but this appears to be the price necessary to allow missing data. We cannot have the degree of the polynomial chosen for one individual depend on the number of observations made on another. In situations where the degree of the polynomial is either known or can be assumed to be the same for each individual, the user has the option of bypassing the individual tests.

Stepping up vs. stepping down

Tests of the form (8) can be performed in two distinct ways. One can choose Q to be small (e.g. $Q = 2$, linearity) and step up (increase Q) if H is rejected; or Q can be chosen large ($Q = P - 1$) and step down if H is accepted. Variations on the first theme obtain depending on the concurrent choice of P : P can be chosen to be large (as was done by Zerbe) or one can choose $P = Q + 1$ in which case (8) reduces to a simple t-test. This latter strategy, namely start with $Q = 2$, $P = 3$ and step up to $Q = 3$, $P = 4$, etc., when necessary, was prevalent in the earlier literature (e.g. Ref. 22, Sec. 15.6), but this has been recognized as potentially misleading inasmuch as accepting, e.g., $\tau_3 = 0$ does not imply that some higher order coefficient is not zero [23].

Anderson (Ref. 24, p. 31) approached the question by casting it into the framework of a multiple decision problem (given Q and P, choose an integer, D, between these numbers) and suggested that the approach of choosing $Q = P - 1$ and stepping down has a number of desirable properties. This was also the approach

used in Ref. 17 and recommended by Ref. 25 (p. 92) and Ref. 26 (p. 41). We have followed these leads in our program. For each individual, in turn, we prompt the user for D, the degree of a polynomial which he/she is confident will be adequate to fit the data (the 'full model'). We then test

$$\text{H: } \tau_p = 0 \tag{9}$$

i.e. we take the 'reduced model' to be of degree $D - 1$. The p-value for this test is printed and the user can use this degree or step down until one of the ensuing sequence of tests is rejected. Note that in this case (8) reduces to a t-test and we provide both the value of t and the corresponding P-value. We also note that Anderson (Ref. 24, p. 38) suggests that different levels of significance can be used as the sequence of tests proceeds. This is possible within our program since the user decides, on the basis of the P-value provided, whether or not to step down. One may, e.g., choose the initial value of D to be fairly large, but to use a small level of significance at this stage: if this degree is necessary, one has a chance of learning that fact, but if it is not, one has but a small probability of choosing it (Ref. 24, p. 42). On the other hand, it has been suggested that $\alpha \approx 0.10$ might be a better choice [27].

Refitting vs. augmentation

Having determined the D_i and $D = \max\{D_i\}$, each of the N vectors of estimated polynomial regression coefficients must be 'expanded,' if necessary, to $P \times 1$ so that the data for each individual will have the same dimension. Zerbe [15] simply augments the $P_i \times 1$ vectors by adding zeros. Dawson et al. [17] refit all individuals to $D = \max\{D_i\}$. Ten Have et al. [20] showed that considerable differences between these two approaches can exist, but it is not clear whether or not one method is 'uniformly better' than the other. While the use of orthogonal polynomials can be expected to minimize the differences between the two strategies [20], we decided to follow Zerbe [15] and Dawson et al. [17] and use the (centered) original time points in our implementation. The feeling that the resulting regression coefficients are easier to interpret persists (Ref. 24, p. 34; Ref. 28, p. 93) and the simple act of centering the data mitigates the computational problems when high degree polynomials are employed while maintaining a convenient interpretational framework. Additionally, as noted earlier, the repeated calls of ORPOL [20] required for accommodating varying patterns of missing data in large data sets would be quite time-consuming.

In our program, both refitting and augmentation with zeros are possible. To refit, one simply chooses a common D for each individual. Augmentation with zeros is automatically accomplished by fitting polynomials of degree D_i to each subject on an individual basis. Examples follow.

Extrapolation and the choice of $[a, b]$

The polynomial regression coefficients for all individuals are estimated over an interval $[a, b]$ specified by the user. This requires that $[a, b]$ be chosen in such a way that each individual growth curve be 'well-behaved' and representative of the individual growth patterns on this interval. Since missing data are allowed, this may re-

quire the extrapolation of one or more growth curves. While Zerbe [15] emphasizes that $[a, b]$ can be selected for its 'biological importance,' this must be balanced with the need for $[a, b]$ to fairly represent the data. In Zerbe's example, where each of the girls was measured twice annually over the age interval [4, 18] with no missing data, he simply identified $[a, b]$ with [4, 18] and this presented no problems. Suppose, however, that one of the girls was measured only on [4, 7]. A line or a quadratic might fit very well *on this interval* but extrapolate poorly to [4, 18] or even to a subinterval of this. The user would then have to decide between adjusting $[a, b]$ or deleting this girl from the analysis. Both are allowed in our program. The user has control over $[a, b]$ and the option of simply deleting a case from the analysis. Deletion may appear to defeat the purpose of allowing missing data, but it might be necessary if we are not to compromise the integrity of the analysis: in order to apply Zerbe's procedure we must have *one* P and *one* $[a, b]$ common to all N individuals and this is not always feasible in practice.

In any event, we have programmed the several decisions which must be made for each individual in determining an appropriate value of D_i in such a way that the user may be able to 'learn from experience'. The program allows the user to either (i) process the subjects beginning with those with the maximum value of T_i and work down, or (ii) process them sequentially beginning with the first in the file. We recommend the first option as this may allow the user to get an early indication of what D might be and facilitate later choices based on fewer numbers of time points.

The Program

The user is asked to provide a $N \times T$ data matrix, **X**, containing the values of the measurements made at times $t_1, t_2, \ldots, t_T$ on each of the N individuals comprising the sample. This data set can be in either *ASCII* format and have the .ASC extension, or in *GAUSS* format with the .DAT extension. Periods ('.') are used to represent missing data values, both here and in the encoded data set. Thus, e.g., the head circumference data (in cm) from achondroplastic infants considered in Ref. 17 might be assembled in a file like

```
.    .    39.4 41.9 .    42.5 43.8 45.7 46.9 47.6 48.3 .    .
34.3 36.8 41.2 42.5 43.8 45.7 46.3 47.6 48.9 49.5 .    .    50.0
.    41.0 .    .    .    .    47.0 .    50.0 .    .    53.0 .
.    40.0 42.0 .    44.5 45.5 .    48.0 .    .    49.8 .    .
38.7 40.0 41.2 43.1 45.1 45.7 .    .    .    .    50.2 .    .
36.8 .    40.3 41.5 42.3 43.8 .    .    .    .    .    .    .
.    40.6 43.8 44.4 45.7 45.7 46.3 46.9 47.6 48.2 48.9 49.5 49.5
34.3 38.1 39.4 40.6 43.2 43.2 46.9 .    .    .    .    .    .
.    40.0 43.0 .    44.3 46.5 47.2 .    49.0 .    52.5 .    53.0
.    40.0 42.0 44.0 45.5 .    46.5 .    .    .    .    .    .
.    .    .    .    .    .    .    .    49.5 50.1 51.4 .    52.4
```

There are $N = 11$ rows (individuals) and $T = 13$ columns, the times of measurement being age in months from 0 to 12. While there is a substantial amount of missing

data, there is at least reasonable overlap of the growth profiles over [0, 12]. We have corrected an apparent typographical error in Ref. 17 where the observations at 10 and 11 months for the seventh individual are interchanged.

The program is invoked with the single command *gsruni zcoeff*. The program menu appears after the program is initiated. With this, the user specifies the directory containing the data set of interest, after which all data files in either *ASCII* or *GAUSS* format are displayed. From the list of available files, the user selects the file of choice. Next the user supplies the value of T; he/she is also given the opportunity to explicitly provide values of $t_1, t_2, \ldots, t_T$ other than consecutive integers beginning with 1 and ending with T (the default).

The data set may also contain a group indicator variable. This is for the convenience of users who may wish to compare the estimated regression coefficients in several groups. If the user indicates that such a variable is included, he/she is first asked if the data are in the default format consisting of a data set with the group variable in column 1, followed by the values of the repeated measurements. If not, the user is prompted for the number of the column containing the grouping variable and the columns (fields) containing the values of the first and last of the repeated measurements. The group indicator variable can be in any column. The longitudinal measurements can also be anywhere in the data set, but they must be in consecutive columns. If in the above example one wanted to compare, say, males and females and if the grouping variable was appended to the above data set in the last (14th) column, the user would respond 14, then 1 and 13.

Once the structure of the data set has been specified, the user is given the option of centering the data and the growth profiles for each of the N individuals are plotted. The user may wish to center the data (use the transformed time scale $t^* = t - \bar{t}$, where $\bar{t}$ is the mean value of t) both to avoid problems of multicollinearity and to facilitate the interpretation of the regression coefficients. For uncentered data, τ_1, the intercept, will often be meaningless, but this is the value of the growth curve in the middle of the range, $t = \bar{t}$, when the data are centered [20]. For convenience, both the centered and original time points appear on the plot.

The user is then asked whether or not he/she wishes to specify a single value of D which will be used for every individual. If YES, we plot the fitted curves for each individual and provide the values of

$$T_i,\ R_i^2,\ t_i \text{ and } p_i \tag{10}$$

for $i = 1, 2, \ldots, N$. Here T_i is the number of measurements, R_i^2 is the square of the multiple correlation coefficient, t_i is the value of the t-statistic for H: $\tau_P = 0$, and p_i is the corresponding p-value. τ_P is the highest order regression coefficient in the model, e.g., if $D = 2$, there are three coefficients and $P = D + 1 = 3$. The user will want to start with a value of D which he/she considers will be adequate to fit the data. The R^2 and p-values will then indicate whether stepping down to a lower degree is feasible (if H is accepted, one would ordinarily try the next smallest degree and compare the new R^2 with the old). A single plot containing all of the individual growth curves with each indicated on a color keyed legend is provided. This feature is useful in identifying individuals with aberrant plots (if any). In plots of data sets

with a large number of subjects, the potential difficulty in distinguishing among the many curves can be overcome by using the graphics utility for 'zooming in' on a small part of the plot and thus single out problematic individuals.

On the basis of the plots and the values of these statistics, the user now chooses between (we continue the above example with $D = 2$):

(a) 'Accept $D = 2$ for all curves, begin analysis' — in which case the user has the option of changing $[a, b]$. The corresponding regression coefficients are then computed and can be saved.

(b) 'Change D for all curves' — in which case the plots and statistics are reproduced for the new value of D.

(c) 'Select an individual curve for refitting' — if abberant plots can be singled out, the user may select several cases for study one-at-a-time. For each case, the user is prompted for a value of D and obtains the corresponding plot and goodness-of-fit statistics. The user then chooses between (i) 'accept this curve with $D = 2$', (ii) 'change D and replot curve' — plots and statistics are produced, (iii) 'delete individual case' — subject removed from subsequent analysis. The first two options are the same as in the single D situation. The third may be exercised when one cannot strike a balance between D and $[a, b]$, i.e. one cannot fairly represent the data over an interval that is important in the context of the study.

If the user replies NO to the question about using a single D for each individual, the growth profiles are plotted individually, one-at-a-time, either starting with individuals having the largest value of T_i and working down, or in the order in which they appear in the file. For each individual, the user is prompted for D, thus allowing the interactive fitting of the curve. Again, the user will generally want to start with a degree which should certainly be adequate and step-down if the test indicates that the highest order regression coefficient is not significantly different from zero. The fitted curve is plotted and the statistics provided. Once again the user may choose to accept the curve with the D selected, change D, or delete the case altogether from the analysis.

Once a common D and $[a, b]$ have been determined, we estimate, and the user may save, the corresponding regression coefficients for the growth curves, the velocity curves and the accelerations in *ASCII* files. Note that D must be at least two for the accelerations to exist. The numerical output is also saved in a file COEFF.OUT which can be modified, highlighted, etc., using a word processor and subsequently printed.

Examples

We consider two examples. The first corresponds to choosing a single, common D for each case and the second to obtaining individual-specific D_i values. Both are based on the data in Ref. 17. In example 1 we treat the data as a one-sample problem (no group structure). In example 2, we build on the earlier example where we wished to compare the male and female cases. Here we assume that the first five cases are

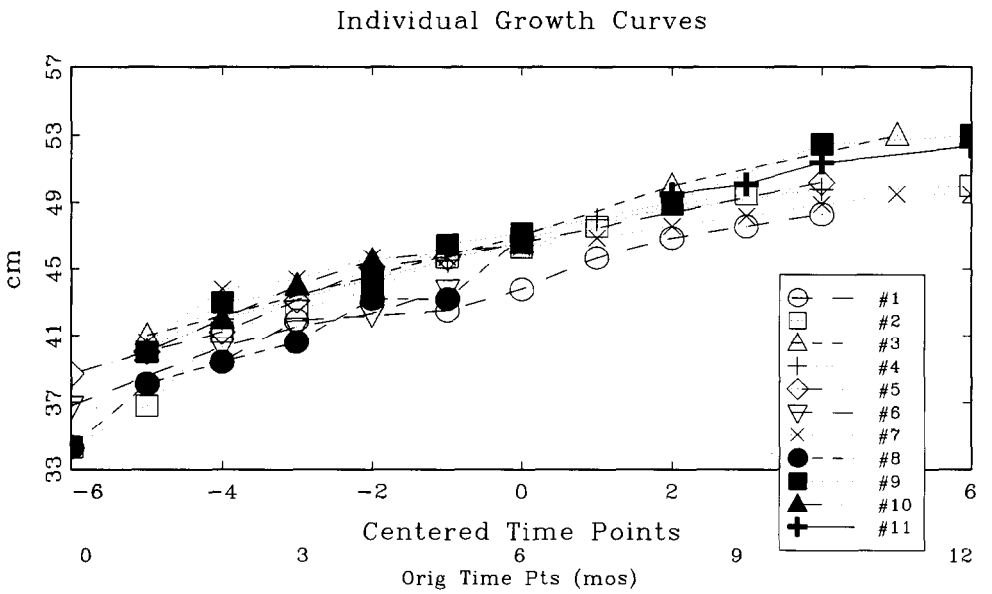

Fig. 1. Individual growth profiles, i.e. original, unfitted longitudinal observations (head circumference) from Dawson et al. [17].

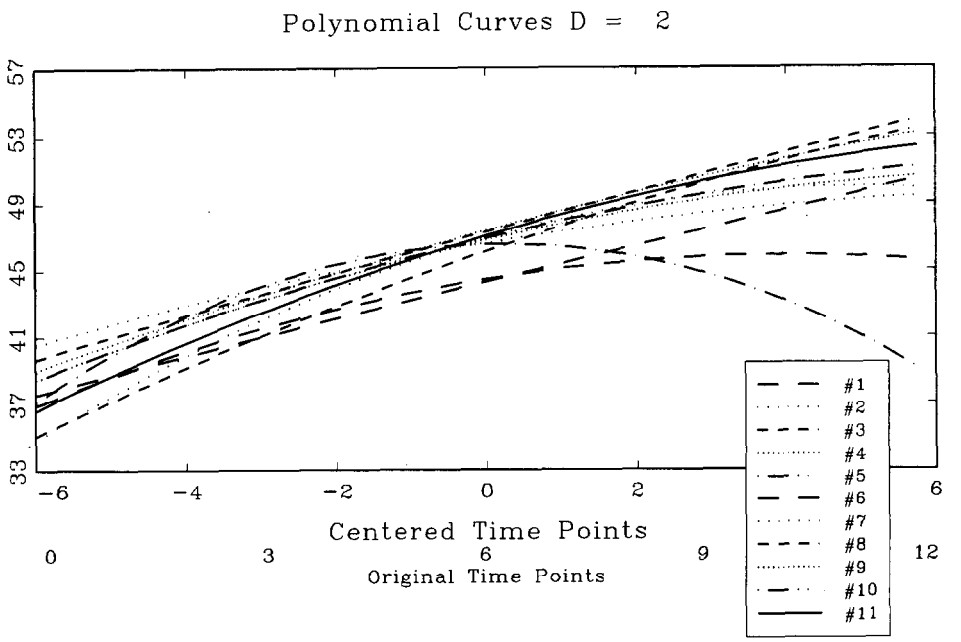

Fig. 2. All subjects fitted with quadratic equations. Note the unlikely down-turn in individual No. 10.

males and the remaining six are females and that the grouping variable is in the first column.

Example 1

If on the basis of the plot of the individual growth profiles seen in Fig. 1 we decide to try $D = 2$, we get the plot of the fitted growth curves seen in Fig. 2 and the corresponding statistics:

Individual	T	R^2	t	P
1	8	0.968	0.144	0.8911
2	11	0.987	8.465	0.0001
3	4	0.997	0.522	0.6937
4	6	0.997	5.190	0.0139
5	7	0.993	3.709	0.0207
6	5	0.993	1.812	0.2117
7	12	0.955	2.703	0.0243
8	7	0.958	0.446	0.6789
9	8	0.979	1.455	0.2053
10	5	0.997	7.569	0.0170
11	4	0.976	0.708	0.6078

The p-values are for the test H: $\tau_3 = 0$, i.e. of whether the quadratic term is significant.

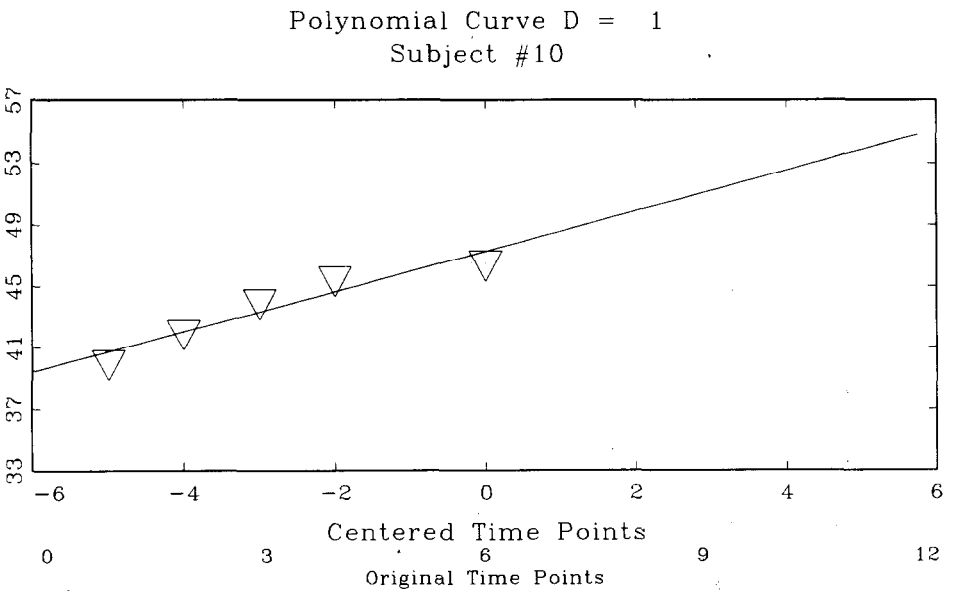

Fig. 3. Individual data for subject No. 10 fitted with a linear equation.

The user can now accept $D = 2$ for all curves, change D for all curves, or select an individual curve for refitting. In this example, it is clear from the plot that individual No. 10 is 'aberrant'; while a quadratic equation fits the data for this individual extremely well ($R^2 = 0.997$) over the range of observations for this individual, extrapolation to the full range [−6, 6] produces an obviously inappropriate growth curve. We therefore select individual No. 10 and specify $D = 1$. We get the plot of the fitted curve seen in Fig. 3 and the statistics

$$T = 5,\ R^2 = 0.918,\ t = 5.797,\ p = 0.0101$$

Since the line appears to provide a reasonable fit over the entire interval [−6, 6], we choose to accept $D = 1$ for this individual. We choose $D = 2$ for the remaining cases and the program augments the vector of estimated regression coefficients for No. 10 with one zero.

The program now produces the estimated regression coefficients. For the growth curves, these are (in terms of the centered time points):

$\hat{\tau}_1$	$\hat{\tau}_2$	$\hat{\tau}_3$
44.278	1.093	−0.005
46.906	1.251	−0.128
47.280	1.213	−0.010
46.836	1.013	−0.065
46.911	1.066	−0.059
44.437	0.719	−0.091
46.609	0.743	−0.046
46.114	1.550	−0.050
47.370	1.196	−0.034
47.270	1.311	0
47.036	1.322	−0.070

Similar output for velocities and accelerations is obtained.

Example 2

If we choose to look at the growth profiles one-at-a-time, starting with individuals having the largest numbers of observations, we are first presented with the growth profile for individual No. 7, with $T = 12$, seen in Fig. 4.

The user is prompted for D. If he/she responds $D = 3$ for this individual, the plot shown in Fig. 5 and the following statistics are obtained:

$$T = 12,\ R^2 = 0.971,\ t = 2.107,\ p = 0.0682$$

The user can now accept this curve with $D = 3$, change D and replot the curve, or delete the case. If the user opts to try $D = 2$, the corresponding plot is provided along with the statistics, as above. When a satisfactory D is found (or the case is deleted), the program proceeds to the next case and continues in this fashion until

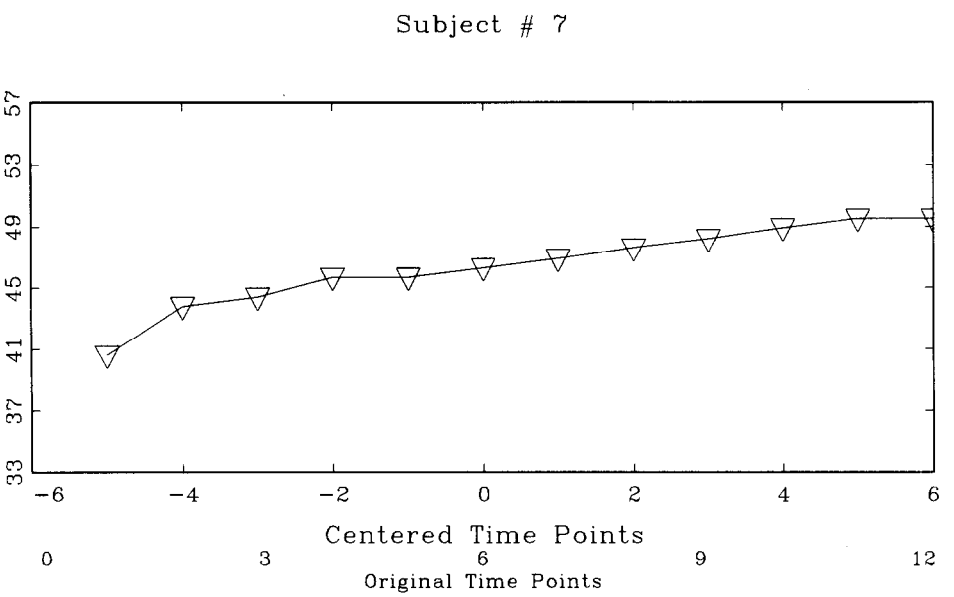

Fig. 4. Growth profile for the individual having the most time-points in the data set, subject No. 7.

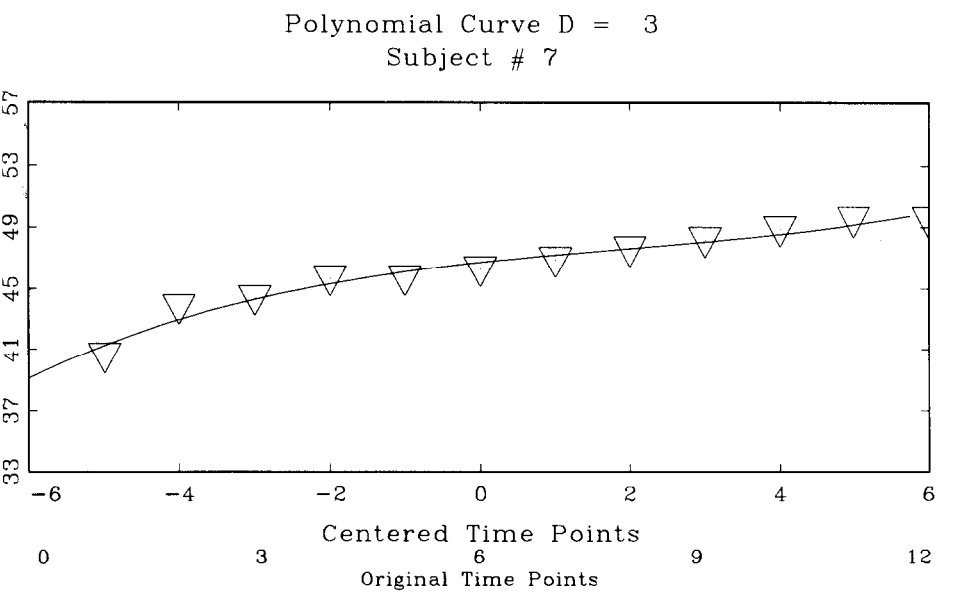

Fig. 5. Subject No. 7 fitted with a cubic equation.

all subjects have been considered. We do not show all the intermediate steps here, but we obtained the following values of D_i on this basis:

$D_i = 1$ for $i = 1, 3, 4, 5, 6, 8, 9, 10, 11$ and $D_i = 2$ for $i = 2, 7$

The program then augments with zeros as necessary. Recalling that in this example we are distinguishing between males and females, the values of the group membership variable and the estimated regression coefficients for the individual growth curves are:

Group	$\hat{\tau}_1$	$\hat{\tau}_2$	$\hat{\tau}_3$
1	44.239	1.093	0
1	46.906	1.251	−0.128
1	47.142	1.217	0
1	46.232	1.084	0
1	46.270	1.170	0
2	45.295	1.361	0
2	46.609	0.743	−0.046
2	46.364	1.850	0
2	46.938	1.166	0
2	47.270	1.311	0
2	48.043	0.749	0

One might, e.g., read the above data into SYSTAT and compare the groups with respect to all three coefficients using Hotelling's T^2 test. Also note that since we opted to use centered time points in this example, τ_1 corresponds to the average fitted values over all time points for a given case (or the fitted value at $t = \bar{t}$). Velocities and accelerations are again produced and made available for further analysis. It is also important to note that velocity (acceleration) curves determined as derivatives of estimated growth curves (as is done in our program) tend to be reliable only over the time interval for which velocity (acceleration) could have been determined by fitting growth increments [8]. Thus, e.g., Zerbe [15] took $a = 4.5$ and $b = 17.5$ for the velocity curves in a study where observations were made from 4 to 18 years at half-year intervals; and $a = 5$, $b = 17$ for accelerations. In our example, one might consider using $a = 1$, $b = 11$ for velocities; $a = 2$, $b = 10$ for accelerations. It will be recalled that the user of our program has control over the choice of a and b and that plots are provided to facilitate this decision.

Discussion

We have emphasized two aspects of Zerbe's procedure, both in the above description and in the structure of our program. First, we pay considerable attention to the determination of the D_i. On the one hand, it is an advantage to represent the trend by a polynomial of low degree because the curve is smoother, the presumed 'explanation' is simpler and the function more economical. If too high a degree is used, there

will be an unnecessarily large variability in the estimation of the trend (Ref. 24, p. 35). The extent of this problem is illustrated in Ref. 5. On the other hand, if the degree is too low, there will be a bias in the estimation of the trend (Ref. 18, p. 117). For example, if we postulate $D = 1$, but in fact $D = 3$ and if the observations are taken at the (centered) time points $-3, -2, -1, 0, 1, 2, 3$, the expected value of the estimator of τ_1 is $\tau_1 + 4\tau_3$, that of τ_2 is $\tau_2 + 7\tau_4$. It is seen, then, that the determination of the D_i is an important part of Zerbe's procedure and that considerable flexibility is necessary if we are to arrive at degrees which fairly represent the data over the interval of interest.

Secondly, close inspection of both the growth profiles and fitted growth curves is strongly recommended. The profiles assist the user in deciding on a convenient starting point (degree) for the goodness-of-fit tests; the fitted growth curves enable one to judge the reasonableness of this and subsequent choices. The goodness-of-fit statistics (R^2 and p) are meant as aids to the user in determining D, but they do not guarantee that a given curve will be appropriate over the entire interval of interest $[a, b]$. Dawson et al. [17] used individual No. 10 in their analysis despite the fact that the fitted curve for this individual 'turns down' over the last 6 months. Their decision to use $D = 2$ for all individuals was reasonable on the basis of the statistics: the quadratic term appears to be necessary ($p < 0.05$) for 5 of the 11 individuals (including No. 10 for which $p = 0.017$) and the R^2 values are uniformly high ($R^2 = 0.997$ for No. 10), but the plot clearly shows that we should either use $D = 1$ for this individual, or delete him/her from the analysis, or change $[a, b]$. Including No. 10 in the analysis undoubtedly resulted in wider than necessary confidence bands in Ref. 17.

In closing, we note that the strategy of replacing the $N \times T$ matrix, **X** by the $N \times P$ matrix of estimated regression coefficients has a long history in longitudinal data analysis, dating back to 1938 and the classic paper by Wishart [29]. Benefits include dimensionality reduction (parsimony), increased power and enhanced interpretability. For a more detailed discussion of these points, refer to Ref. 20.

Acknowledgement

This work was supported by grant DE08730 from the National Institute of Dental Research.

Appendix

A full set of PC programs for longitudinal data analysis, including this program, can be obtained on high density 5.25″ or 3.5″ diskettes (please request type) by sending $25 to defray the cost of handling and licensing fees. These programs require a 80 386 or 80 486 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). 80 386 computers must also be equipped with a 80 387 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 checks payable to Baylor College of Dentistry.

References

1 Rao CR: Some problems involving linear hypotheses in multivariate analysis, *Biometrika*, 46 (1959) 49–58.

2 Rao CR: The theory of least squares when the parameters are stochastic and its application to the analysis of growth curves, *Biometrika*, 52 (1965) 447–458.

3 Kills M: A note on the analysis of growth curves, *Biometrics*, 24 (1968) 189–196.

4 Potthoff RF and Roy SN: A generalized multivariate analysis of variance model useful especially for growth curve problems, *Biometrika*, 51 (1964) 313–326.

5 Schneiderman ED and Kowalski CJ: Implementation of Rao's one-sample polynomial growth curve model using SAS, *Am J Phys Anthropol*, 67 (1985) 323–333.

6 Schneiderman ED, Willis SM, Ten Have TR and Kowalski CJ: Rao's polynomial growth curve model for unequal-time intervals: A menu-driven GAUSS program, *Int J Biomed Comput*, 29 (1991) 235–244.

7 Ten Have TR, Kowalski CJ and Schneiderman ED: A PC program for analyzing one-sample longitudinal data sets which satisfy the two-stage polynomial growth curve model, *Am J Hum Biol*, 3 (1991) 269–279.

8 Schneiderman ED and Kowalski CJ: Implementation of Hills' growth curve analysis for unequal-time intervals using GAUSS, *Am J Hum Biol*, 1 (1989) 31–42.

9 Ten Have TR, Kowalski CJ, Schneiderman ED and Willis SM: A PC program for performing multigroup longitudinal comparisons using the Potthoff-Roy analysis and orthogonal polynomials, *Int J Biomed Comput*, 30 (1992) 103–112.

10 Ten Have TR, Kowalski CJ, Schneiderman ED and Willis SM. Two SAS programs for performing multigroup longitudinal analyses, *Am J Phys Anthropol*, 88 (1992) 251–254.

11 Schneiderman ED, Kowalski CJ and Ten Have TR: A GAUSS program for computing an index of tracking from longitudinal observations, *Am J Hum Biol*, 2 (1990) 475–490.

12 Schneiderman ED, Kowalski CJ, Ten Have TR and Willis SM: Computation of Foulkes and Davis' nonparametric tracking index using GAUSS, *Am J Hum Biol*, 4 (1992) 417–420.

13 Schneiderman ED, Willis SM, Kowalski CJ and Ten Have TR: A PC program for comparing tracking indices in several independent groups. *Am J Hum Biol*, 4 (1992) 399–401.

14 Schneiderman ED, Willis SM, Kowalski CJ and Ten Have TR: A PC program for growth prediction in the context of Rao's polynomial growth curve model. *Comput Biol Med*, 22 (1992) 181–188.

15 Zerbe GO: A new nonparametric technique for constructing percentiles and normal ranges for growth curves determined from longitudinal data, *Growth*, 43 (1979) 263–272.

16 Rubin DB: Inference and missing data, *Biometrika*, 63 (1976) 581–592.

17 Dawson DV, Todorov AB and Elston RC: Confidence bands for the growth of head circumference in achondroplastic children during the first year of life, *Am J Med Genet*, 7 (1980) 529–536.

18 Draper N and Smith H: *Applied Regression Analysis*, Second Edition. Wiley, New York, 1981.

19 Neter J, Wasserman W and Kutner MH: *Applied Linear Statistical Models*, Third Edition. Irwin, Homewood IL, 1990.

20 Ten Have TR, Kowalski CJ and Schneiderman ED: A PC program for obtaining orthogonal polynomial regression coefficients for use in longitudinal data analysis, *Am J Hum Biol*, 4 (1992) 403–416.

21 Joossens JV and Brem-Heyns E: High power polynomial regression for the study of distance, velocity and acceleration of growth, *Growth*, 39 (1975) 535–551.
22 Snedecor GW: *Statistical Methods*, Fifth Edition. Iowa State Univ. Press, Ames IA, 1956.
23 Duran BS: Regression, polynomial. In: *Encyclopedia of Statistical Sciences*, (Eds: S. Kotz, N.L. Johnson and C.B. Read) Vol. 7, Wiley, New York, 1986, pp. 700–703.
24 Anderson TW: *The Statistical Analysis of Time Series*, Wiley, New York, 1971.
25 Plackett RL: *Principles of Regression Analysis*, Oxford Univ. Press, Oxford, 1960.
26 Williams EJ: *Regression Analysis*, Wiley, New York, 1959.
27 Kennedy WJ and Bancroft TA: Model building for prediction in regression based on repeated significance tests, *Ann Math Stat*, 42 (1971) 1273–1284.
28 Goldstein H: *The Design and Analysis of Longitudinal Studies*, Academic Press, New York, 1979.
29 Wishart J: Growth-rate determinations in nutrition studies with the bacon pig, and their analysis, *Biometrika*, 30 (1938) 16–28.