ANNALS OF HUMAN BIOLOGY, 1986, VOL. 13, NO. 2, 129–141

Efficient statistical modelling of longitudinal data

H. GOLDSTEIN
Institute of Education, University of London

Received 10 March 1985; revised 1 October 1985

Summary. A new class of statistical models is proposed for the analysis of longitudinal data, especially those from growth studies. The models are all derived from a simple univariate two-level polynomial model. It is shown that they make efficient use of available data, and can handle a very wide range of problems. They have several important advantages over existing procedures.

1. Introduction

Many human growth studies typically have been based on a small set of measurements at specified ages or occasions, and much of the methodology for modelling growth data has assumed that all individuals will be measured at such a set of target ages. There is no necessary reason however for requiring this, and the present paper presents a new class of models for growth data, and for repeated measures designs in general, which does not require a common set of occasions or ages. This allows considerable flexibility in incorporating explanatory variables other than age into the model, and extends to the class of mixed longitudinal models considered by Patterson (1950), Jones (1980) and Woolson, Leeper and Clarke (1978), for example, and to general multivariate models with missing data.

2. Fitting models to growth data

It is possible to make several distinctions between the various procedures which have been used for fitting human-growth-curve height and weight data. Historically, the earliest and most developed are those which use non-linear, typically logistic, curves for the adolescent growth period, and more recently mixtures of such curves to cover almost the entire age range (Bock and Thissen 1980). A variation is the family of curves recently introduced by Preece and Baines (1978), also non-linear and designed to model the greater part of the age range. Such models can be called parametric, in that they model the age range of interest using a small number of parameters. For example, 30 height measurements taken between 4 and 20 years might typically be modelled by between 5 and 8 parameters. The principal aim of such models is to summarize a complex growth pattern so that the parameters, or functions of them can be used to describe individuals, to relate to other measurements, to make group comparisons and so forth.

While such models undoubtedly have their uses, they suffer from three general drawbacks. First, they tend to impose a too rigid 'shape' upon the growth pattern which does not allow for sufficient individual variation. For example, a logistic curve fitted to the adolescent growth period requires that the height at the time of maximum velocity (PHV) is half-way between the lower and upper asymptotes of the curve, which is not characteristic of individual data (Goldstein 1979). Secondly, because essentially the same curve is fitted to a wide age range, interesting local variations may be missed. Thus in the Preece–Baines curve (model 3) the 'mid-growth spurt' is missed entirely

(Gasser, Muller, Kohler, Prader, Largo and Molinari 1984). Thirdly, these models are unable to account for growth variation attributable to other characteristics measured at each occasion. Thus, for example, the variation about an overall fitted curve may be related to biochemical, environmental or social factors, which vary from occasion to occasion. Seasonal variations are another example of occasion-related data. We shall return to this issue later.

A more recent set of developments which overcomes some of the above problems also involves the fitting of smooth curves to individual growth patterns, but is non-parametric and allows the shape of the curve to be determined locally, that is over a small range of ages, and where the same form of curve is not assumed for each subject. Spline functions and the related kernel estimation techniques have been used (Stutzle, Gasser, Molinari, Prader and Huber 1980, Gasser, Kohler, Muller, Kneip, Largo, Molinari and Prader 1984). These techniques can be regarded as moving averages with varying weighting functions, whereby at any age only nearby measurements contribute to the shape of the curve. These techniques, needless to say, have their own problems: notably the choice of the size of the 'smoothing parameter, so that on the one hand the curve does not merely follow what are best regarded as random variations, and on the other that the smoothing is not too large to pick up interesting local patterns. However, it is worth remarking in this context that a solution to this problem may well lie in collecting more frequent measurements. These models also share the third disadvantage of the parametric models mentioned previously: an inability to describe relationships with other occasion-measured variables. Another issue which has received relatively little attention (but see Silverman 1985) is that of heteroscedastic errors, most importantly an increasing residual variation with age.

Another problem with non-parametric models is that each subject requires a minimum number of data points, and preferably all subjects should have the same number of points. Thus, information of potential use in the estimation of population distributions may be unusable. In spite of such problems, these procedures seem to have several important advantages over the non-linear parametric models, especially for identifying local events of interest such as the mid-growth spurt.

All the above models are based upon fitting separate curves to each subject and then studying the distribution of various derived estimates. The major alternative approach has been to model directly the variations in growth parameters for a sample of individuals using polynomial regression models. A brief history of this approach is given in Goldstein (1979), and a detailed mathematical exposition can be found in Grizzle and Allen (1969) with extensions to handle missing data given by Kleinbaum (1973) and Goldstein (1979). This approach requires that each individual be measured at the same fixed set of r occasions and the measurement at each occasion is then regarded as one of the response variables in an r-variate linear model, with powers of age as explanatory variables. Apart from ease of computation, these models can easily incorporate further explanatory variables which are occasion-defined. As parametric models they may also perform better than the non-linear parametric models for some purposes, and unlike these they can be fitted over small age ranges to estimate local events. Nevertheless, they suffer a major disadvantage compared to single-subject models in that they require a fixed set of occasions and thus are rather inflexible. They also require a choice of the order of polynomial, although the amount of data is generally the important constraint here. To some extent, variability of occasion times can be accommodated by 'adjustment' procedures, but this is not always possible or even desirable (Goldstein 1981).

It is sometimes claimed that biological interpretations can be made from the values of parameters derived from single-subject models. Nevertheless, apart from explicit and not very successful attempts to incorporate biological theory into the determination of curve structure (Goldstein 1979), almost all the development has gone into providing growth summaries which 'follow' observed growth as 'closely as possible'. To do this, the criterion of minimizing residual variation is used—a mathematical rather than a biological criterion. In fact, the real importance of any biology lies in an ability to relate the elements of the models to other biological (or environmental or social, etc.) data, rather than in a determination of the precise form of the curve itself. In the next section we introduce a new class of models which provides wide scope for introducing such variables directly and which overcomes some of the limitations of the models described so far.

3. Growth as a two-level model

Many biological and other data have an hierarchical or multi-level structure. For example, animals have offspring grouped into litters, where the parents comprise the first and higher level and the animals within litters comprise the second and lowest level of the hierarchy. Within any one litter a measurement such as weight will vary among animals, and at the same time the average weight of a litter will vary between litters. In general, the variation of such a measurement can be separated into two parts, one at each level, and appropriate statistical models for studying such 'variance components' are commonly used, for example, in animal genetic studies. We may also view longitudinal data as a two-level hierarchy, in which the first level is the individual subject and the second level is the occasion with subject, as follows:

The basic two-level model is written in two parts:

1. The within-subject model:

$$y_{it} = \Sigma_j \beta_{ij} x_t^j + \Sigma_k \alpha_k z_{itk} + \epsilon_{it}, t = 1, \ldots, n_i \quad (1)$$

with

$$\begin{aligned} \text{cov}(\epsilon_{it}, \epsilon_{it}') &= 0, \\ \text{var}(\epsilon_{it}) &= \sigma_0^2 \end{aligned}$$

where n_i is the number of measurement occasions for subject i, x_t is the age at occasion t and $j(0, \ldots, p)$ indexes the polynomial coefficient. The first summation represents the polynomial (or other curve linear in the β_{ij}) fitted to the set of response measurements. The second summation takes place over a set of further explanatory variables, indexed by k, which in general are measured at each occasion, but may also be constant over occasions, such as gender. In general α_k will be constant over individuals, but variable α_k can be handled in the general model. The β_{ij} on the other hand are assumed to be random variables, and this leads to the specification of the second part of the model:

2. The between-subject model:

$$\beta_{ij} = \beta_j + \gamma_{ij} \quad (2)$$

with

$$E(\gamma_{ij}) = 0$$

The γ_{ij} are not necessarily independent, and we write:

$$U = \{\sigma_{u,jj}'\}$$

where

$$\sigma_{u,jj'} = \text{cov}(\gamma_{ij}, \gamma_{ij'})$$

Thus U represents the covariance matrix of the individual polynomial growth curve coefficients (including the constant term).

This formulation is essentially that of Grizzle and Allen (1969), except that no restriction is now placed on the number or location of time-points t.

Combining (1) and (2), the full model can be written as:

$$y_{it} = \Sigma_j \beta_{ij} x_t^j + \Sigma_k \alpha_k z_{itk} + \epsilon_{it} \quad (3)$$

where

$$\begin{aligned} \beta_{i0} &= \beta_0 + \gamma_{i0} \\ \beta_{ij} &= \beta_j + \gamma_{ij} \end{aligned} \quad (4)$$

It is also possible to define further within-subject random variables. For example, the polynomial coefficients may vary from occasion to occasion and we could write:

$$\beta_{i1t} = \beta_1 + \gamma_{i1} + \delta_{it}$$

with

$$\text{var}(\delta_{it}) = \sigma_1^2$$
$$\text{cov}(\epsilon_{it}, \delta_{it}) = 0$$

then the contribution to the within-subject variance of y_{it}, that is the variance about the individual growth curve, is given by:

$$\sigma_0^2 + \sigma_1^2 x_t^2$$

which allows for increasing variance with age, an important possibility which is not incorporated in the models of the previous section. In fact we can incorporate explanatory variables which are random, with means constrained to be zero, so that they enter only into the error part of the model. Thus, suppose we include the explanatory variable $x_t^{1/2}$ and constrain its correlation with ϵ_{it} to be zero. We then obtain a within-individual contribution to the variance of

$$\sigma_0^2 + \sigma_1^2 x_t^2$$

so that the variance now has a simple linear regression on age. Clearly, a very wide range of variance functions can be specified both within and between subjects. At each level, the error terms in general may be correlated, but we assume zero correlations across levels. There are, however, certain restrictions on the number of independent variance and covariance terms which seem not to have been noted before. For example, if $p=2$ and there are between-subject random terms for each coefficient there are two possible contributions to the variance term involving x_t^2, namely:

$$2\text{cov}(\gamma_{i0}, \gamma_{i2}) x_t^2$$

and

$$\text{var}(\gamma_{i1}) x_t^2$$

so that only one can be estimated and the variance term is constrained to be twice the covariance term.

Equation (3) is a special case of the general mixed-effects multi-level linear model which can have any number of levels and any number of random coefficients at each level. That is the variables z, which can be defined at any level, may be defined as random at any number of levels. Thus, for example, we can measure the number of younger siblings to a child which varies over occasions, or we could measure a child's birth order which is constant for that child and hence defined at the between-subject level. Other variables at this level are birthweight, birth date and height of parents.

Strenio, Weisberg and Bryk (1983) also discuss the two-level growth-curve model. The model they present is a special case of the one given here and assumes only a simple within-child error term, although they make reference to a more general specification. Their model also does not allow for occasion-related variables.

Roughly speaking, the estimation for the general model proceeds as follows. Initial estimates of the polynomial and other coefficients are obtained using ordinary least squares, which ignores the complex structure of the random error terms. The squared residuals based on this fitted model are then regressed on a set of variables defining the structure of the random error terms and thus provide estimates for these. These estimates are then used in a generalized least-squares analysis to obtain a second set of estimates for the coefficients, and so on iteratively until convergence is achieved. When the errors are normally distributed then the procedure is equivalent to maximum likelihood. A detailed specification and estimation procedures for the general model are given in Goldstein (1986). A special case of equation (3) is of some interest. We write:

$$y_{ij} = \Sigma_j \beta_{ij} w_j + \Sigma_k \alpha_k z_{ijk} \qquad (5)$$

where there are now p occasions ($j = 1, \ldots, p$) and $w_j = 1$ if the measurement is at occasion j, 0 if not.

The β_{ij} are all random variables at the between-subject level, and there is now no within-subject variation because a term is fitted for each occasion.

The model (5), without the second summation, is essentially the model given by Jones (1980) for efficient estimation of means in mixed longitudinal studies with a fixed set of occasions, except that he assumes that the covariance matrix is known whereas in the present model it is estimated from the data. In fact (5) also is more general, since the second summation term allows further explanatory variables to be incorporated. Hence our formulation of the general two-level growth-curve model unifies a number of previously separate models. Since the model can readily extend to three or more levels we can also incorporate a hierarchical population structure, and in particular we are able to model a multi-stage sampling of subjects (Goldstein 1986). Furthermore, it is clear from (5) that the y_{ij} need not be the same measurement. Thus j can refer to the jth variable of a multivariate linear model and (5) then allows us to obtain efficient estimates for this model. It also allows for missing response variable values in the same way as for longitudinal data, and in particular can be used to provide efficient estimates of means and covariances for univariate linear models, which in the normal case are equivalent to maximum likelihood (Beale and Little 1975).

Examples

The data are serial growth measurements, targeted a year apart on 72 boys and 66 girls from the London Growth Study (Tanner, Goldstein and Whitehouse 1970). The age range is from 6 to 11 years.

In the first set of analyses, height is related to a basic quartic polynomial in age, and the subject-level variable, gender. We require the polynomial in this age range to provide estimates for both the mid-growth spurt and the pre-pubertal minimum velocity, defined as the ages of zero acceleration. Thus at least a quartic is necessary. The between-individual error terms are fitted for the coefficients up to age squared, higher-order random coefficient terms having zero estimates due to the relatively small sample size. This gives the following contribution to the variance at age t:

$$\sigma_{u,0}{}^2 + 2\sigma_{u,10} x_t + \sigma_{u,1}{}^2 x_t{}^2 + 2\sigma_{u,12} x_t{}^3 + \sigma_{u,2}{}^2 x_t{}^4 \quad (6)$$

The within-subject error model fits the overall constant term and the square root of age, giving a contribution to the variance at age t (the covariance is zero):

$$\sigma_0{}^2 + \sigma_1{}^2 x_t \quad (7)$$

Table 1. Analysis of height related to age, sex and target age.

Explanatory variable	A		B		C		D	
	Estimate	SE	Estimate	SE	Estimate	SE	Estimate	SE
Constant	133·0		133·0		133·0		133·0	
Age	5·26	0·09	5·22	0·09	5·22	0·09	5·22	0·09
Age^2	−0·18	0·04	−0·21	0·03	−0·21	0·03	−0·21	0·03
Age^3	0·023	0·015	0·033	0·012	0·033	0·012	0·033	0·012
Age^4	0·0059	0·0064	0·011	0·005	0·011	0·005	0·011	0·005
Sex	−1·98	1·05	−2·02	1·05	−2·01	1·05	−2·02	1·05
Sex × Age	0·06	0·13	0·13	0·12	0·13	0·12	0·13	0·12
Sex × Age^2	0·12	0·06	0·17	0·03	0·17	0·03	0·17	0·03
Sex × Age^3	0·052	0·002	0·029	0·011	0·029	0·011	0·029	0·011
Sex × Age^4	0·011	0·009	—	—	—	—	—	—
Age–Target age	—	—	—	—	—	—	0·12	0·34
$\sigma_0{}^2$	0·23	0·017	0·16	0·011	0·23	0·017	0·23	0·017
$\sigma_1{}^2$	—	—	0·008	0·013	—	—	—	—
$\sigma^2_{u,0}$	38·1	4·6	38·2	4·6	38·2	4·6	38·2	4·6
$\sigma^2_{u,1}$	0·42	0·05	0·42	0·05	0·42	0·05	0·42	0·05
$\sigma_{u,01}$	2·59	0·37	2·60	0·37	2·59	0·37	2·59	0·37
$\sigma^2_{u,2}$	0·012	0·003	0·012	0·003	0·012	0·003	0·012	0·003
$\sigma_{u,12}$	0·040	0·008	0·040	0·008	0·040	0·008	0·040	0·008
$\varrho_{u,01}$	0·65		0·65		0·65		0·65	
$\varrho_{u,12}$	0·56		0·56		0·56		0·56	
Number of iterations	6		9		6		6	

Relative accuracy for convergence = 0·0001.
Age is measured from 9·0 years except for the term $\sigma_1{}^2$.
Sex is coded: boy = 0, girl = 1.
Number of children = 138.
Number of measurements = 778.

Table 1 displays the results of four analyses. The first three are with and without the term $\sigma_1{}^2 x_t$ in (7), and with and without the (age)^4 by gender interaction, which is non-significant in analysis A. There is a difference between analyses B and C in the estimate of the within-subject variance. In the former case this increases from 0·21 at age 6 to 0·25 at age 11. The value of $\sigma_1{}^2$, however, is less than its standard error and for simplicity it is assumed zero in other analyses, although with a larger sample and perhaps wider age range a detailed modelling of the within-subject variation would be useful.

Table 2. Height related to age for each sex.

Explanatory variable	Boys ($N=72$)		Girls ($N=66$)	
	Estimate	SE	Estimate	SE
Constant	133·0		131·0	
Age	5·26	0·08	5·31	0·11
Age^2	−0·18	0·03	−0·068	0·047
Age^3	0·023	0·013	0·074	0·019
Age^4	0·0056	0·0055	0·017	0·008
σ_0^2	0·17	0·017	0·30	0·032
$\sigma_{u,0}^2$	38·6	6·44	37·7	6·58
$\sigma_{u,1}^2$	0·27	0·047	0·59	0·095
$\sigma_{u,01}$	2·52	0·49	2·60	0·54
$\sigma_{u,2}^2$	0·0053	0·0019	0·021	0·006
$\sigma_{u,12}$	0·0021	0·0045	0·085	0·020
$\varrho_{u,01}$	0·78		0·55	
$\varrho_{u,12}$	0·06		0·76	
Number of iterations	8		5	
Number of measurements	411		367	

The average within-subject variance of 0·23 is consistent with other studies. Between individuals, the constant and linear coefficients are quite highly correlated (0·65), as are the linear and quadratic coefficients (0·56).

Table 2 shows separate polynomial coefficients for each gender, and also allows different error term parameters, especially noticeable in the covariances for the boys. The fixed coefficients are similar to those in analysis A of table 1, but there are differences in some of the random parameters. Figure 1 shows the estimated between-

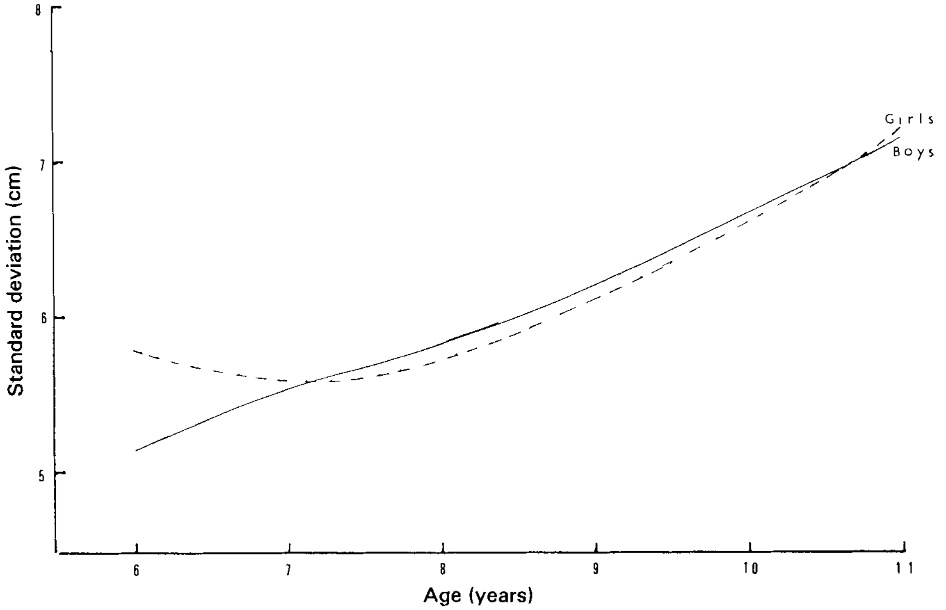

Figure 1. Between-subject standard deviation estimated from models in table 2.

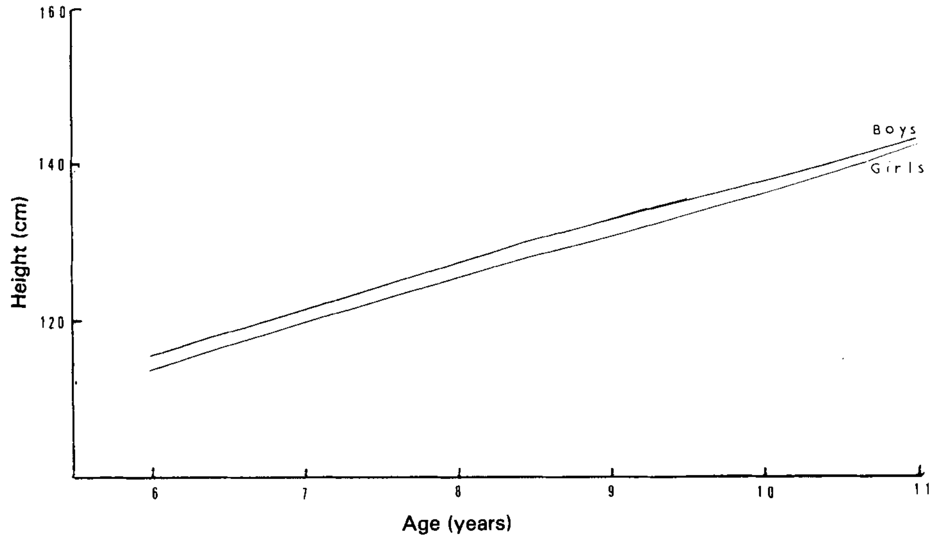

Figure 2. Mean height estimated from model C in table 1.

individual variances by age for each gender. For most ages, these are reasonably close to those given by Tanner, Whitehouse and Takaishi (1966), but that for the girls at 6 years is high, presumably because of the relatively small sample size.

Figure 2 shows the predicted mean heights for boys and girls based on analysis C in table 1. The girls agree more closely with the British norms than the boys who are somewhat taller, but again not significantly so.

We can also use the estimated parameters to calculate various derived statistics. Two such are the mid-growth spurt and the minimum velocity before pubertal 'take-off' growth. The age of occurrence of the former is given as that where a maximum of the velocity occurs around ages of 6 or 7, and the latter as a minimum of the velocity at around ages 9 or 10. Both estimates are found by estimating the ages of zero acceleration. If we write the fixed part of the model as:

$$y_i = a + bx_i + cx_i^2 + dx_i^3 + ex_i^4$$

then the zero acceleration ages are given by the solution of the quadratic

$$c + 3dx + 6ex^2 = 0$$

that is

$$x = \{-3d \pm (9d^2 - 24ce)^{1/2}\}\{12e\}^{-1} \tag{8}$$

From analysis C in table 1 we obtain the following estimates. For the mid-growth spurt, the mean ages are 6·0 years for girls and 6·3 years for boys. For the minimum pre-pubertal velocity they are 9·2 for girls and 10·2 for boys. These estimates compare with 6·1 and 6·4 for the mid-growth spurt and 9·7 and 10·9 for the pre-pubertal ages derived from the information provided in Gasser *et al.* (1984). Other authors seem to find larger gender differences for the mid-growth-spurt ages (Tanner and Cameron 1980). For the pre-pubertal estimates the present results agree closely with those from the data of Tanner and Cameron (1980), namely 9·1 for girls and 10·3 for boys, based on a large sample of London children.

It would seem that the pre-pubertal minimum velocity is experienced by virtually all children but that not all children can be identified as having a mid-growth spurt (Tanner and Cameron 1980). Typically, the existence of a mid-growth spurt appears to be based on a subjective judgement of observed or smoothed individual growth curves. A precise definition of the spurt requires a zero acceleration associated with a maximum of the velocity in a specified age range. For the present data, since we are unable to fit the fourth-degree coefficient as a random term, each child is effectively constrained to have a mid-growth zero acceleration, except in the small number of cases where the quadratic equation has no real solution. This occurs in about 6% of girls and 2% of boys and also implies no real estimate for the pre-pubertal point either. Better estimates of these proportions require the fitting of higher-order random terms, which needs larger sample sizes. It should be noted that Gasser *et al.* (1985) do not provide estimates of the mid-growth zero acceleration ages. They define the mid-growth spurt in terms of the age midway between the maximum acceleration and maximum deceleration within the mid-growth age range, and somewhat misleadingly equate this to the age of peak velocity. It would seem to be more desirable to retain the definition of a mid-growth spurt as occurring only when a velocity maximum occurs. This then maintains consistency of definition with the pubertal spurt.

It should also be noted that the estimate of the magnitude or the age of specific growth events, based on yearly velocities calculated from only two measurements, as in the case of Tanner and Cameron (1980), will not necessarily coincide with that based on fitting curves to more than two measurements for each individual. This is analogous to the comparison of estimates of peak height velocity based on cross-sectional as opposed to longitudinal data (Tanner 1962). In the present case the difference between estimates of the age at minimum growth velocity, based on multi-occasion curve fitting and two-occasion velocities, is approximately proportional to the average ratio of the growth velocity one year following the actual age of minimum velocity to the growth velocity one year preceding it. This difference is unlikely to be large, however, and even when the ratio is as much as two, it would be only about two months.

The rate of change of acceleration is $0{\cdot}198+0{\cdot}264x$ for boys and $0{\cdot}372+0{\cdot}264x$ for girls, which is less than zero (giving a maximum of the velocity) when x is less than 8·25 for boys and less than 7·60 for girls. Hence, the mid-growth spurt as predicted by the model, and subject to sampling error, can only occur at ages below these, and the pre-pubertal minimum velocity at ages above these.

For the pre-pubertal minimum velocity, figure 3 gives the estimated cumulative probability function for each gender. This is derived from equation (8), by simulating normal distributions for the random variables based upon their parameter estimates given in table 1. The 5th and 95th percentiles are 7·6 and 9·9 for girls and 9·0 and 10·6 for boys. In the present model the distribution of this age is non-normal, and it is therefore difficult to compare with the results of Gasser *et al.* (1984) who quote only the mean together with an SD of about 1·0 for both boys and girls.

We can also introduce a variable which is occasion-defined to see if it can explain more of the between- or within-subject variation. In the present study, the aim was to measure each child on their birthday. Some children, however, were early or late. One hypothesis is that children measured early may have arrived early because of concern about relatively slow growth. Thus, if we define a variable which is the difference between the actual age of measurement and the target age, this hypothesis suggests it

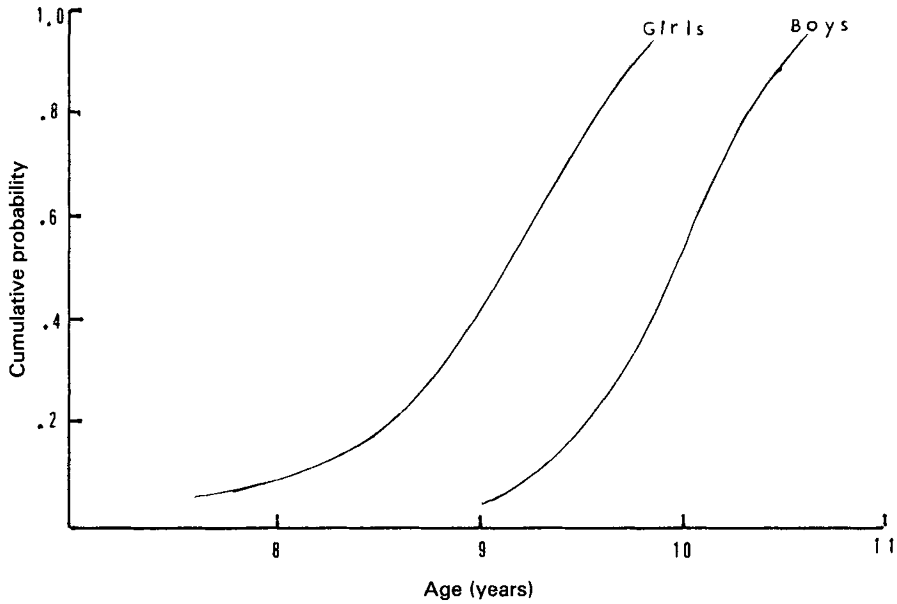

Figure 3. Estimated cumulative distribution of age of pre-pubertal minimum velocity.

should have a positive coefficient in the model. Model D in table 1 incorporates such a term and although the coefficient is indeed positive, it has a relatively large standard error.

The final example illustrates the case described above, of a mixed longitudinal study with fixed occasions and more than one variable measured on each subject. We use the data for the two occasions of eight and nine years of age together with height, weight and gender. Since subjects were not all measured at the target ages, an exact analysis is carried out by fitting a first-order polynomial model, that is a straight line, to the exact ages of measurement.

The model is now essentially a three-level model. The highest level is the subject, the next level is the occasion within subject and the lowest level is the variable (height or weight) within occasion. The model can also be viewed as a two-level model with four variables defined by the combination of variables and polynomial terms. The random variation is at the subject level.

Table 3 gives the fitted means for height and weight for boys and the girl–boy differences, together with the 4×4 covariance and correlation matrices. It will be seen that the height estimates are close to those of the previous analysis.

The improvement in efficiency when several variables are analysed together in a multivariate longitudinal model is important for those variables where there are many missing observations. Hence, in a growth study where this is the case by accident or by design, these models use all the data in an efficient manner. Furthermore, the estimates from such analyses can then be used in further univariate or multivariate analyses which require input in the form of a covariance matrix and a set of means. Even where all individuals attend at exactly the target ages, the advantage of fitting polynomials is that the model may be able to be specified with a relatively small number of random parameters, rather than one parameter for each variable and occasion combination.

Table 3. Height and weight related to sex and age on two occasions.

Explanatory variable	Estimate	SE
Height		
Constant	132·9	0·7
Age	5·5	0·1
Sex	−1·7	1·0
Sex × age	−0·02	0·2
Weight		
Constant	29·0	0·6
Age	3·2	0·2
Sex	−1·1	0·9
Sex × age	−0·3	0·3

Covariance matrix of random coefficients for constant and age (correlations in brackets)

		Height		Weight	
		Constant	Age	Constant	Age
Height	constant	36·3			
	age	2·1 (0·42)	0·7		
Weight	constant	24·8 (0·75)	1·9 (0·40)	30·0	
	age	4·8 (0·43)	0·7 (0·46)	7·8 (0·76)	3·5

Predicted correlation matrix for target ages 8·0 and 9·0 years

		Height		Weight	
		8·0	9·0	8·0	9·0
Height	8·0	1·0			
	9·0	0·99	1·0		
Weight	8·0	0·76	0·78	1·0	
	9·0	0·73	0·75	0·96	1·0

Number of iterations = 8 (starting values for the variance terms, were estimated from the tables in Tanner *et al.* (1966))

Number of children = 148

Number of measurements: height, 257; weight, 257.

3. Discussion

The model and worked examples demonstrate several points.

1. The two-level model provides a means of studying general hypotheses about factors influencing the pattern of growth. Single-curve fitting procedures on the other hand are not well suited to such tasks.

2. All the available data can be used for estimation even if only one or two measurements are present for some individuals. This is a feature not shared in general by other parametric or non-parametric models, although the parametric models are somewhat more efficient in this respect than the non-parametric ones.

3. We may often require a set of simple growth summaries rather than a detailed specification of growth. These summaries, for example, might be in terms of average velocities and acclerations together with their between-subject variabilities and a description of how these interact with other factors.

4. Polynomials are not the only curves which can be modelled as illustrated in this

paper. Any curve linear in its parameters, such as the infant growth curve (Count 1943), can be modelled in this way:

$$y = a = bx_t + c\log(x_t)$$

A special problem occurs with the end of growth. Non-linear curves with an upper asymptote are usually used in this case, with the asymptote estimating adult size. A simple polynomial curve may not be adequate here, and linear growth curves containing, for example, negative exponential terms, could be considered.

5. It is clear from the analyses that the estimates, especially of the random terms, depend on an adequate number of individuals in the sample. In general, parametric models should be used with care when attempting to estimate the occurrence of specific growth events which are functions of the estimated parameters. In the example we appear to get rather good estimates of the mid-growth spurt and pre-puberty minimum velocity ages, but less satisfactory estimates of the variability of these ages through being unable to fit higher-order random terms. Thus it may be the case, at least for small samples, that smoothing procedures, including eye smoothing, will be better for this purpose. In practice a reasonable sequence of analyses might begin with fitting individual curves in an exploratory spirit, and following this by a more formal parametric modelling along the lines of this paper.

6. There is the problem shared by all curve-fitting procedures of the spacing of observations. The more widely these are spaced, the more likely it is that particular events such as growth spurts will be missed in some children. If, on the other hand, occasions are grouped too closely together we may find ourselves describing relatively unimportant short-term fluctuations, although if these can be related to other factors, for example, seasonal ones, this may indeed be useful, and in this respect the two-level model is able to handle such factors efficiently.

The procedures in this paper allow us directly to model the within-subject variation and, for example, to study how this may change with age. Likewise, any other explanatory variable which is occasion-defined may have a coeffcient which is a random variable, and contribute to the within-subject variation. One such variable might be the choice of measurer or measuring instrument. Some measurers or instruments may be more variable than others, and making allowance for this should improve the other parameter estimates.

The analyses in this paper have been confined to a small part of the total age range. This could easily be extended in either or both directions, for example, to ages 14 and 16 respectively in girls and boys in order to include the adolescent peak velocity, and requiring at least a polynomial of degree 5. Whether it is better to fit a single polynomial to a wide age range or several lower-order polynomials to overlapping narrower age ranges is a matter for empirical study, which it is hoped will be pursued. A further topic of interest is the fitting of polynomials to the end of the growth period, where we might expect them to perform relatively poorly.

Finally, by considering a three-level model where the lowest level is the variable measured within measurement occasions, or for example, replications of a single variable, we obtain a flexible procedure for the analysis of multivariate longitudinal data, which automatically handles missing data, and which provides more efficient estimates for growth variables where there are many missing measurements.

Acknowledgements

I am most grateful to the following who have provided valuable suggestions for the improvement of this paper: Professor M. Aitkin, Dr T. Cole, Mr J. R. Ecob, Professor M. J. R. Healy, Mr N. Longford, Mr I. Plewis and Professor J. M. Tanner.

References

BEALE, E. M. L., and LITTLE, R. J. A., 1975, Missing values in multivariate analysis. *Journal of the Royal Statistical Society, B*, **37**, 129–145.

BOCK, R. D., and THISSEN, D., 1980, Statistical Problems of fitting individual growth curves. In *Human Growth and Motivation: Methodologies and Factors*, edited by F. E. Johnston, A. F. Roche and C. Susanne (New York: Plenum).

COUNT, E. W., 1943, Growth patterns of the human physique: an approach to kinetic anthropometry. *Human Biology*, **15**, 1–32.

GASSER, T., KOHLER, W., MULLER, H. G., KNEIP, A., LARGO, R., MOLINARI, L., and PRADER, A., 1984, Velocity and acceleration of height growth using kernel estimation. *Annals of Human Biology*, **11**, 397–411.

GASSER, T., MULLER, H. G., KOHLER, W., PRADER, A., LARGO, R., and MOLINARI, L., 1985, An analysis of the mid-growth and adolescent spurts of height based on acceleration. *Annals of Human Biology*, **12**, 129–148.

GOLDSTEIN, H., 1979, *The Design and Analysis of Longitudinal Studies* (London: Academic Press).

GOLDSTEIN, H., 1981, Some graphical procedures for the preliminary processing of longitudinal data. In *Interpreting Multivariate Data*, edited by V. Barnett (Chichester: Wiley).

GOLDSTEIN, H., 1986, Multilevel mixed linear model analysis using iterative generalised least squares. *Biometrika*, **73**, 43–56.

GRIZZLE, J. E., and ALLEN, D. M., 1969, Analysis of growth and dose-response curves. *Biometrics*, **25**, 357–361.

JONES, R. G., 1980, Best linear unbiased estimation in repeated surveys. *Journal of the Royal Statistical Society, B*, **42**, 221–226.

KLEINBAUM, D. G., 1973, A generalisation of the growth curve model which allows missing data. *Journal of Multivariate Analysis*, **3**, 117–124.

PATTERSON, H. D., 1950, Sampling on successive occasions with partial replacement of units. *Journal of the Royal Statistical Society, B*, **12–13**, 241–255.

PREECE, M. A., and BAINES, M. J., 1978, A new family of mathematical models describing the human growth curve. *Annals of Human Biology*, **5**, 1–24.

SILVERMAN, B. W., 1985, Some aspects of the spline smoothing approach to non-parametric regression curve fitting. *Journal of the Royal Statistical Society, B*, **41**, 1–52 (with discussion).

STRENIO, J., WEISBERG, H. I., and BRYK, A. S., 1983, Empirical Bayes estimation of individual growth-curve parameters and their relationship to covariates. *Biometrics*, **39**, 71–86.

STUTZLE, W., GASSER, T., MOLINARI, L., PRADER, A., and HUBER, P. J., 1980, Shape-invariant modelling of human growth. *Annals of Human Biology*, **7**, 507–528.

TANNER, J. M., 1962, *Growth at Adolescence* (Oxford: Blackwell).

TANNER, J. M., and CAMERON, N., 1980, Investigation of the mid-growth spurt in height, weight and limb circumferences in single-year velocity data from the 1966–67 London Growth Survey. *Annals of Human Biology*, **7**, 565–577.

TANNER, J. M., GOLDSTEIN, H., and WHITEHOUSE, R. H., 1970, Standards for Children's height at ages 2–9 years allowing for height of parents. *Archives of Diseases in Childhood*, **45**, 755–762.

TANNER, J. M., WHITEHOUSE, R. H., and TAKAISHI, M., 1966, Standards from birth to maturity for height, weight, height velocity and weight velocity. British children 1965. *Archives of Diseases in Childhood*, **41**, 454–471, 613–635.

WOOLSON, R. F., LEEPER, J. D., and CLARKE, W. R., 1978, Analysis of incomplete data from longitudinal studies. *Journal of the Royal Statistical Society, A*, **141**, 242–252.

Address correspondence to: Professor H. Goldstein, Department of Mathematics, Statistics and Computing, Institute of Education, Bedford Way, London WC1, UK.

Zusammenfassung. Eine neue Klasse statistischer Modelle wird für die Analyse von Längsschnittdaten vorgeschlagen, besonders für jene aus Wachstumsuntersuchungen. Alle Modelle werden aus einem einfachen, univariaten, zweistufigen, polynomen Modell abgeleitet. Es wird gezeigt, daß sie die verfügbaren Daten wirkungsvoll ausnutzen, und daß sie eine sehr weite Spanne von Problemen behandeln können. Sie haben mehrere wichtige Vorteile gegenüber vorhandenen Prozeduren.

Résumé. Une nouvelle classe de modèles statistiques est proposée pour analyser des données longitudinales, particulièrement celles d'études de croissance. Les modèles sont tous dérivés d'un modèle simple univarié polynomial à deux niveaux. Il est montré qu'ils font un usage efficace des données disponibles, et peuvent traiter une gamme très large de problèmes. Ils ont plusieurs avantages très importants par rapport aux procédures existantes.