131

Rev. Epidém. et Santé Publ., 1989, *37*, 477-484

Flexible models for the analysis of growth data with an application to height prediction

Modèles flexibles pour l'analyse de la croissance. Application à la prévision de la taille

H. GOLDSTEIN

Institute of Education, University of London, 20 Bedford way, London WCIH OAL, U.K.

Cet article montre comment les développements récents de la modélisation à niveaux multiples peuvent s'appliquer à l'analyse des données de croissance et en particulier à la prédiction de la taille adulte. Cette approche est à la fois efficiente et souple sur le plan statistique.

Courbe de croissance. Prédiction de la taille. Modèle à niveau multiple. Données multivariates.

We show how recent developments in the theory of multilevel statistical modelling can be applied to the analysis of growth data and in particular to the prediction of adult height. This approach is both statistically efficient and very flexible.

Growth curve. Height prediction. Multilevel model. Multivariate data.

INTRODUCTION

Recent work on the specification and estimation of so called "hierarchical" or "multilevel" statistical models has provided us with a powerful new tool for the analysis of longitudinal growth data. In these models it is supposed that measurements are hierarchically structured. Thus, for example, students belong to the lowest level of a hierarchy, the upper levels of which are classrooms and schools. Likewise, repeated measurements on subjects are grouped or clustered within subjects, so that the measuring occasion constitutes the lowest level (1) of the hierarchy and subjects constitute level 2. In the next section we introduce a specific statistical model which captures that structure. First, however, we give a brief review of growth data analysis.

Much of the literature on the analysis of longitudinal growth data, especially height, has dealt with the problem of curve fitting and increasingly mathematically complex forms have been developed for this. A discussion of the usefulness of these is given in Goldstein [1]. On the other hand, relatively little attention has been paid to modelling features of growth in relation to other factors such as social class, or other developmental measures.

In this paper we present a simple yet powerful class of models which not only provide a flexible characterisation of growth curves but also easily allow the incorporation of further explanatory factors which might be related to growth. In addition we show how the models can be used to provide an efficient system for the prediction of adult growth status which has important advantages over previous methods [2].

POLYNOMIAL GROWTH CURVES

The idea of fitting simple polynomial curves to repeated measurements on individual

Texte reçu le 29 juin 1987. Acceptation définitive le 1 avril 1989.

subjects was systematically proposed by Wishart [3]. A full expositon of a multivariate statistical model base on polynomial curve fitting was given by Rao [4].

Briefly, in the simple case we suppose each sample individual has a set of p measurements taken at a fixed set of occasions. For each subject we fit up to a p-1 order polynomial and then study the variation of the resulting coefficients accross subjects. Thus we might find significant variation in just the intercept and slope coefficients. This would then give us and description of growth for the sample, based on an overall average p-1 order polynomial, with individual suject variation about this in the intercept and linear coefficients. In other words, each individual has his own unique curve in terms of the intercept and linear components plus a common higher order component. In addition there is an about-curve-within-individual "residual" variation. We can also go on to study group differences or the dependence of the polynomial coefficients on further factors.

The major problem with these traditional models has been their rigidity in requiring, usually, a fixed set of occasions with no "missing" measurements. The advantage of the 2-level model formulation is that it avoids this rigidity entirely and provides statistically efficient estimates for any number and spacing of measurements per individual. Where there is a fixed set of measurement occasions with no missing data it coincides with the traditional model. Furthermore, the models discussed by Patterson (1950) and Tanner (1951) for so called "mixed" longitudinal data are special cases of the general 2-level model.

The flexibility of the 2-level model has important implications for study design, freeing us from the need to ensure that all individuals are measured at the same fixed set of occasions. An introduction to multilevel models, and the 2-level model in particular, is given by Goldstein [5]. The next section sets out the statistical model and its assumptions.

THE 2-LEVEL GROWTH CURVE MODEL

The basic 2-level polynomial growth curve model can be written as;

$$Y_{ij} = \Sigma_m \alpha_m Z_{mij} + \Sigma_t \beta_{tij} X_{ij}^t \qquad t = 0,....,q \qquad [1]$$
$$\beta_{tij} = \beta_t + U_{ti} + e_{tij} \qquad t = 0,....,q$$

where X_{ij} is time or age, and the Z_{mij} are a set of further explanatory variables or covariates which may vary from occasion to occasion. In the simplest case there is a single random variable at level 1 (occasions), namely e_{oij} representing a constant within-subject variance about the growth curve. We assume :

$$COV(u_{ti}, e_{tij}) = COV(e_{tij}, e_{tik}) =$$
$$E(u_{ti}) = E(e_{tij}) = 0$$

The β_{tij} are the polynomial coefficients and the u_{ti} are the deviations of each subject's coefficient from the mean coefficient for the population. The independence of the level 1 residuals is a strong assumption in some applications. For example, growth in height has a seasonal component, so that where two or more measurements during a year are made this will be superimposed on the underlying growth curve. Failure to model this will result in dependencies among the level 1 residuals. Work on such models is currently being pursued but in the present paper we shall attempt to avoid this problem by using only yearly measurements on the subjects. Nevertheless, as we show below, we can still model, in a simple fashion, the level 1 variance as a function of age or other variables.

The u_{ti} are random variables at level 2 (subjects), giving rise to $q+1$ variances and $q(q+1)/2$ covariances. The coefficients α_m may be fixed or random. This model is discussed in detail, for example by Strenio *et al.* [6] and Goldstein [1]. In general, the u_{ti} for large values of t will be set to zero, so that random between-subject variation is described by the low order polynomial coefficients.

We can extend model [1] into a new class of very general models in which we have more than one response measurement and each has a separate polynomial regression on age with its own set of covariates, and where the random coefficients are correlated across the measurements at each level. Thus, for example, if height and weight are repeatedly measu-

red in growing children, the intercept, linear etc. growth curve coefficients of height, at the subject level, will be correlated among themselves and with those for weight. An advantage of such a multivariate model is that, via these intercorrelations, it can provide efficient estimates for measurements with large numbers of randomly missing measurements. A simple bivariate example is given in Goldstein [1].

In the present paper we consider an example which essentially is a special case of such a general model. The first variate is height, modelled as a function of age and certain covariates, and the second is adult height modelled simply as a function of the overall mean. There are two populations of subjects which have been sampled.

The model is written as follows;

$$Y_{ij} = \alpha_{1i}\delta_{ij}Z_{1i} + \alpha_{2i}\delta_{ij}Z_{2ij} + \delta_{ij}(\Sigma_t \beta_{tij} X_{ij}^t) + (1-\delta_{ij})\gamma_1 + (1-\delta_{ij})\alpha_{31}Z_{1i} \qquad t=0,....,5 \qquad [2]$$

where Z_{1i} is a dummy (0,1) variable indicating whether the subject belongs to group 1 or 2, and is thus a measurement made at the subject level. The variable Z_{2ij} is the subject's bone age, estimated from a wrist radiograph according to Tanner *et al.* [2]. The variable δ_{ij} is 1 if the response is made during the growth period and 0 if adult height is measured, and X_{ij} is age, measured about a suitable origin. The coefficients α_{1i}, α_{2i}, α_{3i} are assumed to be fixed, and the remaining coefficients are assumed to be random as follows;

$$\begin{aligned}
\beta_{oij} &= \beta_o + u_{o1} + e_{oij} && t=0,1\\
\beta_{1ij} &= \beta_1 + u_{1i} + e_{1ij}\\
\beta_{2ij} &= \beta_2 + u_{2i}\\
\beta_{3ij} &= \beta_3 + u_{3i}\\
\beta_{tij} &= \beta_t && t=4,....,5\\
\gamma_i &= \gamma_o + v_i
\end{aligned}$$

At the subject level the random variables u_{oi}, u_{1i}, u_{2i}, u_{3i}, v_i, have a 5-variate distribution with a zero mean vector and dispersion matrix Ω_2. At level 1, e_{oij} and e_{1ij} have a bivariate distribution with a zero mean vector and dispersion matrix Ω_1. Thus, at any given age during the growth period the variance of Y_{ij} is given by;

$$X_2^T\Omega_2X_2 + X_1^t\,\Omega_1X_1$$

where

$$X_2^T = (1, x_{ij}, x_{ij}^2, x_{ij}^3)$$
$$X_1^T = (1, x_{ij})$$

The age range of growth considered in the example is 11 years to 17 years together with measurements of adult height, in a sample of boys. During this period it is well known that there is a maximum of the velocity of growth at puberty and a minimum velocity approached as growth slows down at the approach to adulthood. There is also a pre-pubertal minimum of the velocity but for nearly all boys this occurs before the age of 11 years (Goldstein, [1]). It is also well known that the ages of occurrence of these zero "acceleration" points vary between subjects.

To capture these growth features we require at least that growth coefficients up to the cubic vary randomly between subjects, since the age of zero acceleration is estimated by setting the second differential of the growth curve with respect to age, to zero.

In another paper (Goldstein, [7]) the same model [2] was fitted to a sample of data for girls, where it was not possible to fit a random cubic coefficient because of the relatively small sample size. The problems raised by this are discussed in that paper.

ESTIMATION AND PREDICTION

The estimation procedure used is that described in Goldstein [8] and [5] namely iterative generalised least squares (IGLS) which is maximum likelihood when the random variables have a multivariate gaussian distribution. Software written at the London Institute of Education has been used.

Our interest is primarily in predicting γ_i, the adult height for individuals not in the sample. The mean γ_o is obtained from the model estimates and we can form a posterior estimate of v_i in the usual manner. This is based on the

estimated covariance matrix of adult height and the growth coefficients. Given a set of observed heights during growth for an individual, we can write down the covariance matrix between these measurements and adult height, and so derive a linear regression prediction of the latter given the values of the former. As typically is done when using such procedures we ignore the sampling error of the random parameter estimates when calculating the standard errors of the predictions. In fact the sample size of 110 cases appears to be large enough to justify this. Explicit formulae for the prediction equations and the standard errors of the predicted values are given in Goldstein [5].

DATA ANALYSIS

The data for this example are measurements on two samples of boys measured from just after birth to adulthood. The first sample, known as the International Children's Centre London sample (ICC) consists of 69 boys born in the early 1950's in an area of central London. The second sample (NCH) consists of 41 boys in a children's home in Hertfordshire measured from entry to the home until adulthood. In both samples the children were measured close to their birthdays, and more frequently during periods of rapid growth. We have selected the yearly measurements from the 11th birthday onwards. Further details of the samples are given in Tanner *et al.* [2]. At each measuring occasion height was measured and bone age assessed according to the Tanner-Whitehouse scale [2].

TABLE I. — *Height related to age, bone age, and group. Boys aged 11-17 years — Taille en relation avec l'âge, l'âge osseux et le groupe. Garçons âgés de 11 à 17 ans.*

Fixed coefficients	A Estimate	s.e.
Adult Height	174.5	0.77
Growth curve Intpt	152.8	0.68
Bone age	1.09	0.10
Group (adult)	0.20	0.30
Age	5.72	0.20
Age^2	−0.49	0.08
Age^3	−0.16	0.03
Age^4	−0.053	0.013
Age^5	0.0090	0.0015

Random Coefficients

level 2 covariance matrix (correlations)

	Adult height	Growth Intpt	Age	Age^2	Age^3
Adult height	63.2				
Growth Intpt	44.4(0.79)	49.5			
Age	2.30(0.27)	2.21(0.29)	1.17		
Age^2	0.80(0.24)	−1.00(−0.33)	0.11(0.24)	0.18	
Age^3	−0.10(−0.01)	−0.02(−0.03)	−0.09(−0.96)	−0.02(−0.52)	0.008

level 1 variance = 1.14, s.e. = 0.11

Group is coded 1 if in ICC sample, 0 if in NCH sample.
Age is measured about an origin of 13.0 years.
Number of subjects = 110
Number of measurements = 626

The data for this example are measurements on two samples of boys measured from just after birth to adulthood. The first sample, known as the International Children's Centre London sample (ICC) consists of 69 boys born in the early 1950's in an area of central London. The second sample (NCH) consists of 41 boys in a children's home in Hertfordshire measured from entry to the home until adulthood. In both samples the children were measured close to their birthdays, and more frequently during periods of rapid growth. We have selected the yearly measurements from the 11th birthday onwards. Further details of the samples are given in Tanner *et al.* [2]. At each measuring occasion height was measured and bone age assessed according to the Tanner-Whitehouse scale [2].

Table I gives the parameter estimates from fitting the model [4]. The term for study difference during growth was very small and has been omitted.

The ages of maximum and minimum height velocity are obtained by solving the following equation;

$$\beta_{2ij} + 3\beta_3 x + 6\beta_4 x^2 + 10\beta_5 x^3 = 0 \qquad [5]$$

If we use the estimates for the variance of β_{2ij} and β_{3ij} and their covariance in *table I*, and assume that the coefficients have a gaussian distribution then we can estimate the distribution of x. This is done conveniently by using simulation, and for each simulated set of coefficients finding the value of x which gives a maximum in the age range 11 to 15 years. *Table II* gives estimates for some percentiles of this distribution.

TABLE II. — *Estimated Percentiles of the age of Maximum Height Velocity. — Percentiles estimés de l'âge du maximum de croissance.*

Percentile	Age
5	13.0
10	13.3
50	13.8
90	14.3
95	14.6

mean age = 13.8 years

The mean age of 13.8 years agrees well with that of 13.9 years found by Tanner *et al.* [9] using a sample of the NCH children, including the measurements made every three months, by a method based upon smoothing each individual subject's curve with a logistic curve.

Further checks on the model can be made by plotting standardised (shrunken) residuals at level 1 and level 2 and examples are given in Goldstein [7]. Such plots are not presented here, and did not indicate any unusual patterns.

In *figure 1* we plot the predictions against adult height using the subsample of 78 boys who have a measurement within 0.1 years of their 13th birthday. In *figure 2* we plot the predictions against adult height using the subsample of 43 boys who have three measurements, around their 13th, 14th, and 15th birthdays.

We see clear linear relationships with a larger scatter of the observed residuals about their predictions for the age 13 prediction. At this age the estimated standard error of the prediction is 5.0 cm, for the set of 3 measurements it is 3.9 cm.

The residual variance of 1.14 is considerably larger than the value of 0.14 quoted by Tanner et al. [9] and that of 0.23 for the age range 6-11 years found by Goldstein [1-8]. It

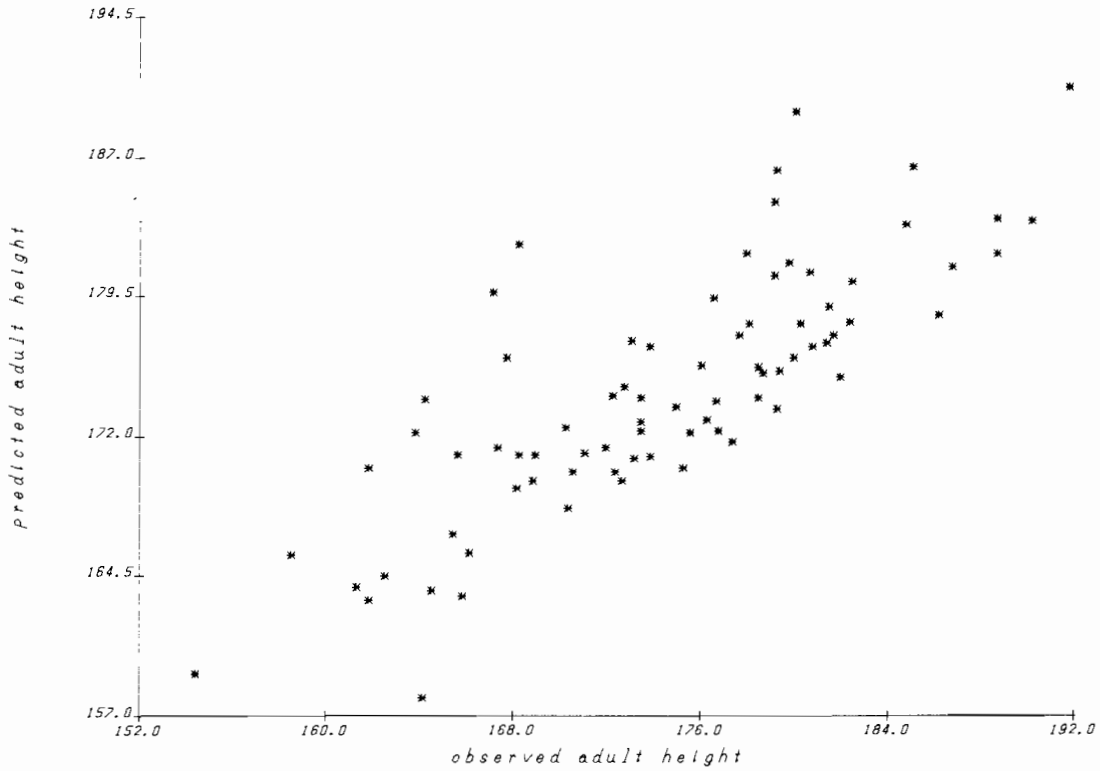

FIG. 1. — boys age 13. — Garçons mesurés à 13 ans

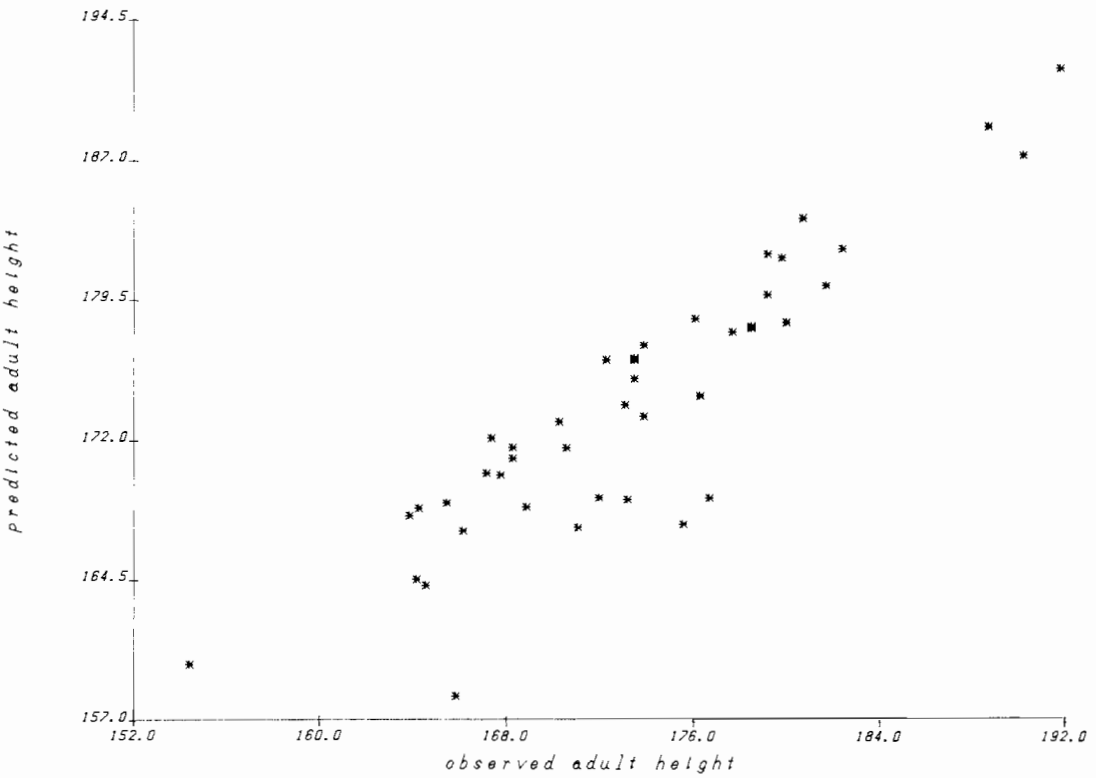

FIG. 2. — boys measured at 13, 14, 15. — Garçons mesurés à 13, 14 et 15 ans.

seems that a model with a random quartic coefficient may be needed, but the present sample size is too small to achieve numerical convergence when we attempt to fit such a model.

DISCUSSION

The analyses in this paper have demonstrated the feasibility of using a 2-level model for simultaneously modelling growth and predicting adult height from measurements taken during the growth period. Clearly, the method can be extended to other measurements and we can also consider the multivariate case where several measurements are modelled jointly. In addition, the adult measurements to be predicted need not to be those measured during growth, and this provides a flexible approach to the modelling of general repeated measures data. For routine use, a program can be written to make predictions with associated interval estimates, and the prediction can be updated as further measurements become available. It should be noted that the adult height predictions are generally population dependent. In the present analysis the group difference is small and non significant, but we cannot necessarily assume in general that all population differences will have been taken into account by conditioning on growth measurements. This will be a matter for empirical study. Likewise, it will often be necessary to adjust for a "secular trend" in adult height which has occurred between the time period when the data were collected and the period the results are in use.

We have assumed simple multivariate distributions among the measurements and the random parameters. In fact, in the case of height data there are some constraints which ought to be included in the model, namely that, for any individual, the adult height cannot be less than any of the growth measure-

ments. Thus, using the models in this paper it would be possible to predict an adult measurement less than the most recent growth measurement. This especially will be the case for growth measures taken towards the end of the growth period. The problem is one which affects all height prediction methods and needs further study.

Two other methods are in use for prediction of adult height. The one by Tanner *et al.* [2] is based upon separate regression predictions of height at each age, or pair of ages. While this procedure can in principle produce efficient predictions, it is not very flexible. Thus, the accuracy of the prediction equation is limited by the actual number of subjects at the age being used, whereas the 2-level model procedure can use efficiently all the data available; including those cases without an adult height measurement. Also the fixed-age prediction method cannot realistically handle more than two serial measures, whereas the 2-level procedure can include as many as are available.

The other procedure [10] is similar to the present one but instead uses a non-linear model fitted to the whole growth age range with parameters varying between subjects.

Detailed comparisons of these procedures have not yet been carried out. The 2-level polynomial model, however, would seem to be the most flexible and potentially the most efficient of these methods. It can handle multiple measurements easily, it can model within individual changes in variation, it can make use of data from individuals with only very few measurements, and it can handle measurements other than height, for which simple non-linear growth models are unavailable.

A further development would be to extend the number of covariates in the model. Thus, Tanner *et al.* [2] effectively include the occurrence of menarche as a covariate by presenting separate predictions for those girls who have and who have not yet experienced that event. Likewise, other stages of pubertal development could be included. The inclusion of subject-level variables such as parental height and birth order might also be useful. In some cases, it may be preferable to treat a continuous occasion-related covariate as a response. Thus, we could fit a bivariate growth model to height and bone age, where in a simple model bone age might be a quadratic function of age with all the coefficients random at the subject level, the intercept and quadratic coefficients having a mean value of zero and the linear coefficient having a mean value of 1.0. The predictor of adult height would then be a function of the set of height and bone age residuals. An important advantage of this model is that even where bone age is not measured at all occasions, all the available bone age measurements can be used in the prediction (Goldstein [7]). This contrasts with previous models where we must use either all the occasions without bone age or just those that contain bone age.

Finally, it should be stressed that large samples are important to secure stable estimates and to enable higher order fixed and random coefficients to be included so that the model can be properly specified. It would be convenient, for example, to be able to model a much wider age range than that considered here, in order to avoid the problem discussed at the end of section 3 and that would require further higher order random coefficients to cope with at least two more stationary values of height growth in the prepubertal period. The optimum combination of overlapping age ranges is a matter for further empirical study. Further work is also needed on the modelling of measurements made close together in time where serial correlations will be present at level 1.

REFERENCES

1. Goldstein H. : Efficient statistical modelling of longitudinal data. *Annals of Human Biology*, 1986, *13*, 129-141.
2. Tanner J.M., Whitehouse R.H., Cameron N., Marshall W.A., Healy M.J.R., and Goldstein H. : *Assessment of Skeletal Maturity and Prediction of Adult Height* (TW2 Method), London, Academic Press, 1983.
3. Wishart J. : Growth rate determinations in nutrition studies with the bacon pig, and their analysis. *Biometrika*, 1938, *30*, 16-28.

4. Rao C.R. : The theory of least squares when the parameters are stochastic and its application to the analysis of growth curves. *Biometrika*, 1965, *52*, 447-58.

5. Goldstein H. : *Multilevel Models in Educational and Social Research.* London, Griffin; New York, Oxford University Press, 1987.

6. Strenio J., Weisberg H.I., Bryk A.S. : Empirical Bayes Estimation of individual growth curve parameters and their relationship to covariates. *Biometrics*, 1983, *39*, 71-86.

7. Goldstein H. : Models for multilevel response variables with an application to growth curves. In : Bock, R.D. *Multilevel Analysis of Educational Data.* New York, Academic Press.

8. Goldstein H. : *Multilevel mixed linear model analysis using iterative generalised least squares*, 1986, Biometrika, 73, 43-56.

9. Tanner J.M., Whitehouse R.H., Marubini E., Resele L.F. : The Adolescent Growth spurt of Boys and Girls of the Harpenden Growth Study. *Ann. Hum. Biol.*, 1976, *3*, 109-126.

10. Bock R.D. : Unusual Growth Patterns in the Fels Data. In, Demirjian, A. (ed.) Human Growth : *A Multidisciplinary Review.* London and Philadelphia, Taylor and Francis, 1986.