130

Efficient prediction models for adult height

Harvey GOLDSTEIN

Institute of Education, 20 Bedford Way, London WC1H 0AL, UK.

Introduction

Elsewhere, (Goldstein, 1986), I have discussed the merits of polynomial modelling over the use of nonlinear or nonparametric models for the analysis of longitudinal growth data. In particular I justified the use of polynomials on the grounds of simplicity and flexibility. A polynomial of sufficiently high order 'graduates' the average growth trajectory, and some of the lower order coefficients are assumed to vary randomly across subjects. Thus, for the age range 6–11 years, I showed (Goldstein, 1986) that it was possible to fit an average fourth order polynomial with the intercept, linear and quadratic coefficients varying across subjects. Such models, however, do have their problems, one of which is their poor performance towards the end of growth when an upper asymptote is required.

In the case of adult height prediction, some approaches involve the fitting of growth curves with an upper asymptote which is used to provide an estimate of adult height. In this chapter the prediction of adult height is considered which avoids the need to fit an upper asymptote, while using a polynomial model.

Several authors (Strenio *et al.*, 1983; Goldstein, 1986, 1987) have discussed the advantages of using a two-level model for the specification and analysis of polynomial growth curves. Briefly, these include the ability to use measurements obtained at any set of ages from a suitable sample of individual subjects, and the ease with which occasion-related or subject-related covariates can be incorporated into the model.

The two-level model is a special case of what have come to be called 'multilevel' models. These have found a wide application in the analysis of social and educational data (Goldstein, 1987) and are designed to cope with hierarchically structured data which typically are found in the social sciences. Thus, students are grouped into schools which are grouped into local education authorities or boards. Likewise, longitudinal repeated-measures data can be viewed as having a two-level structure where measurement occasions (level 1) are grouped within subjects (level 2). Essentially the same statistical modelling techniques can be applied to any set of hierarchically structured data. In the next section a two-level model is developed which has some useful applications for the analysis of longitudinal growth data.

The two-level growth model

In this section the two-level model is outlined with simple examples. The appendix gives the full statistical model.

Suppose, for simplicity, that each subject's growth is linear. We can write:

$$y_{ij} = a_i + b_i x_{ij} + e_{ij} \qquad (1)$$

The first term on the right hand side of (1) is the intercept of the growth line for the ith subject, and the coefficient of x is the slope of the line. The terms e_{ij} represent variation about the growth line, assumed to be random and independent. The subscript *j* refers to measurement occasion and each subject in general will have their own set of measurement ages.

We could fit such a growth line for each subject, assuming at least two measurements on each, and interpret or further analyse the resulting coefficient estimates. More often, however, we would prefer to regard each subject as a (random) member of a population and we then require statements about the average values of the coefficients their variation across subjects, group differences, etc. Thus, for example, we might wish to know the mean slopes in different subpopulations and the between-subject variation in slope.

We now write down a generalization of (1) which will allow us to pursue such analyses:

$$y_{ij} = \beta_{0i} + \beta_{1i}x_{ij} + e_{ij} \tag{2}$$

The first term on the right hand side of (2) is the intercept for the *i*th subject and we designate it as varying randomly across subjects by writing:

$$\beta_{0i} = \beta_0 + u_{0i}$$

The first term on the right hand side is the population mean intercept and the second term is a random variable with: $E(u_{0i}) = 0$, $var(u_{0i}) = \sigma^2_{u0}$. Likewise; $E(u_{1i}) = 0$, $var(u_{1i}) = \sigma^2_{u1}$ and; $cov(u_{0i}, u_{1i}) = \sigma_{u01}$ with $var(e_{ij}) = \sigma^2_0$.

Elsewhere (Goldstein, 1986) I have considered a further modelling of e_{ij}, the within-subject variation, but we shall not pursue this further here.

The basic model (2) can be extended in a straightforward manner to allow for more complex polynomial growth, with some or all the polynomial coefficients varying and covarying across subjects. It can also be extended to include covariates. Suppose, for example, that we have a second order polynomial growth curve with height as the response variable, and we wished to include bone age as a covariate. We would write:

$$y_{ij} = \beta_{0i} + \beta_{1i}x_{ij} + \beta_{2i}x^2_{ij} + \alpha z_{ij} + e_{ij} \tag{3}$$

The intercept and the coefficients of the linear and quadratic terms vary and covary across subjects, ie at level 2, and *e* varies within subjects, ie about each subject's growth curve, at level 1. The covariate, *z*, is the bone age. In this case bone age is an 'occasion level' covariate since it changes from measurement occasion to measurement occasion. Other covariates, such as birth order are 'subject level' ones since they are constant within each subject. In model (3) we have four 'fixed' parameters to estimate, viz, the mean values of the three growth coefficients and the coefficient of bone age. We also have seven 'random' parameters to estimate—three variances and three covariances between subjects plus the within subject variance. The statistical estimation procedure provides efficient estimates of all these parameters and allow us to estimate standard errors, confidence intervals etc. It also enables us to 'predict' individual subject's growth coefficients.

Although the model only assumes that each subject's coefficients vary randomly, given a set of measurements on a subject and estimates of all the parameters, we can obtain a 'regression' prediction of each of that subject's growth coefficients. The more measurements that are made on a subject, the more accurate that prediction will be. Details can be found in Goldstein (1987). It is this feature which we will use when we come to consider the prediction of adult height.

The multivariate growth model

In all the above elaborations of the basic model we have assumed a single response variable, height. We now show how the model can be extended to consider more than one response variable. Consider the case where both height and bone age are treated as responses. We can write a simple model as follows:

efers to measurement occasion
rement ages.
at least two measurements on
timates. More often, however,
of a population and we then
heir variation across subjects,
w the mean slopes in different

us to pursue such analyses:

(2)

e ith subject and we designate

ntercept and the second term
$E(u_{1i}) = 0$, $var(u_{1i}) = \sigma_{u1}^2$ and:

lling of e_{ij}, the within-subject

ner to allow for more complex
varying and covarying across
. for example, that we have
onse variable, and we wished

(3)

erms vary and covary across
h subject's growth curve, at
an 'occasion level' covariate
asion. Other covariates, such
ithin each subject. In model
values of the three growth
en 'random' parameters to
ects plus the within subject
mates of all these parameters
It also enables us to 'predict'

vary randomly, given a set of
we can obtain a 'regression'
measurements that are made
be found in Goldstein (1987).
e prediction of adult height.

d a single response variable,
ler more than one response
reated as responses. We can

$$y_{ij} = \beta_{0i} + \beta_{1i}x_{ij} + e_{ij} \qquad (4)$$

$$y'_{ij} = \beta'_{0i} + \beta'_{1i}x_{ij} + e'_{ij} \qquad (5)$$

Equation (4) is the same as (2) and (5) is the corresponding equation for bone age. In the appendix, for convenience, the full model is written as a single equation using indicator variables. The assumptions made about equation (2) are the same for (4) and (5), with the addition that there are level-2 covariances between all the following random parameters:

$$\beta_{0i}, \beta'_{0i}, \beta_{1i}, \beta'_{1i}$$

There is also the possibility that the level-1 random variables,

$$e_{ij}, e'_{ij}$$

are correlated. Thus, for example, at a measurement occasion a positive deviation from a subject's own growth curve in height might be associated with a positive deviation in his bone age curve. Preliminary analyses of the present data indicate that such correlation is small. This will in part be due to the large amount of measurement error present in both height and bone age after each subject's overall growth curves have been fitted. In the following models, therefore, this covariance is assumed to be zero.

As with model (3), we can extend (4) and (5) to include higher order terms and covariates. The estimation procedure now provides estimates of all the fixed and random parameters including the additional covariances.

Adult height y^* is incorporated into the model simply by writing an additional equation:

$$y_i^* = \alpha_i^* \qquad (6)$$

Equation (6) models adult height as a population mean and a level 2, between subject, random variation with a variance and also covariances with all the other level-2 random variables. Equation (6) can be extended by adding covariates, for example for group differences.

It should be noted that, so long as measurements are made on subjects randomly sampled from a well-defined population, there is no requirement that the same ages are sampled or that the same number of measurements are made on each subject. Thus, we can include subjects with just adult measurements, or just one height measurement, etc. Since many-occasion longitudinal measurement sets tend to be few in number, this enables us to increase the accuracy of parameter estimates with purely cross-sectional data. Of course, a single height growth measurement contributes to the estimate of the height intercept, but little to the estimation of high order coefficients or to estimates of covariances between coefficients.

Data analysis

The data consist of measurements on two samples of boys measured from 11 to 16 years together with their height when adult. The first subsample, known as the International Children's Centre sample (ICC) consists of 69 boys born in the early 1950s in an area of central London. The second subsample (NCH) consists of 41 boys in a children's home in Hertfordshire measured from their entry into the home. The children were measured close to their birthdays and more frequently during the adolescent period of rapid growth (Tanner *et al.*, 1983). Yearly measurements only have been used, for reasons which will be discussed later. A similar sample of a total of 93 girls is also available, but results are not given here. The reason for this is that a satisfactory fit of the model could not be obtained with this number of subjects and the large number of random parameters involved. We shall return to this issue in the discussion.

We first present results from the fitting of growth curves separately to bone age and height, and then present the full model including adult height.

Bone age

Bone age is calculated from the TW2 20 bone score (Tanner *et al.*, 1982), being the average age of subjects at a given bone score value. It is simpler to work with bone age rather than the bone score since on average the bone age, at any given age, will be equal to that age, with no overall trend with age. Thus we use a model where the response is (bone age−chronological age), the explanatory variables consisting of a constant term and chronological age. While the fixed coefficients in this model should be close to zero, individual subjects may vary in their intercept and slope parameters, so these are made random at level 2.

Table 1 shows that the expected value of bone age is 0.22 years greater than age itself, with a standard error of 0.09 years. This indicates that the sample on average is slightly advanced compared to the original population on which bone age was standardized. The slope is estimated as 0.03 units per year and is of the same order as its standard error, so that there is no evidence for any overall increase with age.

From the estimated between-subject covariance matrix we see that there is a standard deviation of 0.9 years for the intercept and a standard deviation of 0.2 units per year for the slope, with a correlation between them of −0.24. There is thus considerable variation in the bone maturity for subjects: an approximate 95% confidence interval being −1.5 to 1.9 years. An approximate 95% confidence interval for the slope is −0.4 to 0.5 units per year. These estimates quantify the known fact that individuals vary both in their degree of advancement and their rate of maturity development. The within subject, level 1, standard deviation is 0.4 units, much of which may be measurement error.

Height

A fifth order polynomial has been fitted for height, with all the coefficients up to and including the cubic, random at level 2. At level 1 a simple variance parameter is fitted. Since the cubic coefficient varies randomly between subjects, the age of zero acceleration, or peak height velocity, will also vary between subjects, as it should over this age range.

All of the fixed coefficient estimates are statistically significant at the 5% level, except the quintic. The estimates of the correlations between the linear and quadratic, and between the quadratic and cubic coefficients, are very small. It will often be useful to omit parameters which are poorly estimated, in order to stabilize the estimates of the remaining parameters. This is particularly important for random parameters when the sample size is relatively small.

It will also be seen that the estimated correlation between the linear and cubic coefficients is slightly greater than 1.0. This is a result of sampling variability.

Table 1. Bone age minus chronological age, related to chronological age.

Fixed coefficient	*Estimate*	*s.e.*	Level-1 variance = 0.18
Intercept	0.22	0.09	Age measured about 13.0 years
Age	0.03	0.03	Number of subjects = 108
Random parameters			Number of measurements = 436
Level 2:	*Covariance matrix (correlation)*		
	Intercept	*Age*	
Intercept	0.76		
Age	−0.05(−0.24)	0.06	

:parately to bone age and height,

er *et al.*, 1982), being the average
work with bone age rather than
ı age, will be equal to that age,
ıere the response is (bone age–
constant term and chronological
to zero. individual subjects may
nade random at level 2.
ears greater than age itself, with
· on average is slightly advanced
ndardized. The slope is estimated
·rror, so that there is no evidence

ve see that there is a standard
·ion of 0.2 units per year for the
ıus considerable variation in the
nterval being – 1.5 to 1.9 years.
).4 to 0.5 units per year. These
in their degree of advancement
evel 1, standard deviation is 0.4

e coefficients up to and including
ameter is fitted. Since the cubic
:leration, or peak height velocity,
ınge.
ant at the 5% level, except the
ınd quadratic, and between the
useful to omit parameters which
ıe remaining parameters. This
ample size is relatively small.
he linear and cubic coefficients
ability.

o chronological age.

.-1 variance = 0.18

neasured about 13.0 years
ber of subjects = 108
ber of measurements = 436

Table 2. Height related to a fifth degree polynomial in age.

Fixed coefficient	*Estimate*	*s.e.*	
Intercept	153.0	0.76	Level-1 variance = 0.82
Age	7.03	0.22	Age measured about 13.0 years
Age^2	0.64	0.18	Number of subjects = 108
Age^3	– 0.26	0.09	Number of measurements = 436
Age^4	– 0.08	0.04	
Age^5	0.02	0.02	

Random parameters

Level 2:	*Covariance matrix (correlation)*			
	Intercept	*Age*	*Age^2*	*Age^3*
Intercept	59.5			
Age	3.30(0.27)	2.51		
Age^2	– 1.38(– 0.42)	0.02(0.03)	0.18	
Age^3	– 0.22(– 0.18)	– 0.26(– 1.03)	0.003(0.04)	0.03

Height, bone age and adult height

We now combine the separate height and bone age models and add adult height as a third response variable with a mean value and simple between-subject variation. The results are given in Table 3.

The results for height and bone age are similar to those obtained in the separate analyses, with an additional height intercept and a coefficient of the dummy indicator variable specifying which subsample a subject belongs to for the adult height part of the model. The correlation between adult height and the intercept term, representing height at age 13 years is high, being 0.85. The correlation between height at age 13 and the subjects' bone age deviation is 0.44. As before, the correlations between the linear and quadratic and between the quadratic and cubic are very small. The correlation between the bone age deviation and the height slope coefficient is only 0.03, and the covariance has a relatively large standard error. When we try to estimate the covariances between the bone age deviation and the coefficients of the quadratic and cubic for height, we find that the iterative estimation procedure becomes unstable. Thus, these have been omitted.

In the present model we have estimated 21 random parameters (variances and covariances) at level 2, based on a sample of 110 level-2 units (subjects). With such a high ratio of parameters to units convergence of the iterative procedure is often slow and difficult. The present analysis should be regarded as a pilot study in methodology. A more extensive analysis needs to be carried out, bringing together different data sets, with allowance made for appropriate subpopulation differences.

Predicting adult height

As remarked earlier, given the values or estimates of the fixed and random parameters, we can predict an individual subject's growth coefficients. Full details can be found in Goldstein (1987) and the following is a brief sketch.

Returning to equation (2), suppose we wished to estimate individual i's intercept and slope coefficient, given a set of actual measurements. That is, we wish to find $\boldsymbol{u}_{0i}$, and $\boldsymbol{u}_{1i}$.

Since we have estimates of the variances and the covariance between the intercept and slope, we can derive the covariances between the intercept and the slope and any set of actual measurements (measured as deviations from their predicted values using the 'fixed' part of the model). This in turn allows us to calculate the regression coefficients of the intercept and

Table 3. Combined model for adult height, height, and bone age.

Fixed coefficient	*Estimate*	*s.e.*
Adult height:		
Intercept	174.4	0.78
Subgroup	0.25	0.50
Height:		
Intercept	153.0	0.72
Age	6.91	0.20
Age^2	0.43	0.09
Age^3	−0.14	0.03
Age^4	−0.03	0.01
Age^5	0.03	0.03
Bone age:		
Intercept	0.21	0.09
Age	0.03	0.03

Random parameters

	Level 2:	*Covariance matrix (correlation)*					
	Adult height	*Intercept (ht)*	*Age (ht)*	*Age^2 (ht)*	*Age^3 (Ht.)*	*Intercept (B. A.)*	*Age (B. A.)*
Adult height	62.5				0.08 (0.06)	0.57 (0.08)	—
Intercept (ht)	49.5 (0.85)	54.5			0.01 (0.01)	3.00 (0.44)	—
Age (ht)	1.11 (0.09)	1.14 (0.09)	2.5		0.27 (1.02)	0.02 (0.01)	—
Age^2 (ht)	0.39 (0.12)	−0.39 (−0.12)	0.05 (0.08)	0.17	0.00 (0.00)	—	—
Age^3 (ht)					0.03	—	—
Intercept (B. A.)					—	0.85	—
Age (B. A.)					—	−0.09 (−0.39)	0.06

Level-1 variance (ht) = 0.89
Level-1 variance (B. A.) = 0.18
Age measured about 13.0 years
Number of subjects = 110
Number of measurements = 982

Table 4. Standard errors for predicted adult height at specified ages of measurement.

Bone age measure (yr)			Height measure (yr)		
		—	11.0	11.0 12.0	11.0 12.0 13.0
—			4.3	4.2	4.1
11.0		7.9	3.9	3.8	3.7
11.0, 12.0		7.9	3.7	3.7	3.5
11.0, 12.0, 13.0		7.9	3.5	3.5	3.3

slope, treated as responses, on the actual measurements treated as predictors. Thus we obtain predicted values of the intercept and slope for an individual subject by applying this prediction equation. We can also obtain estimates of the standard errors of the predicted values.

In the full model given in the appendix, each subject has their own adult height which is treated as a random variable in the model. Thus, given any set of growth period height and bone age measurements, we can obtain a prediction of that subject's adult height. This prediction can be made using whatever measurements are available, since these are just the predictor variables in the regression. Of course, the more such variables there are the more accurate the prediction will be. A computer program is available which will carry out the necessary computations given a set of input measurements.

Table 4 shows how the standard error of predicted adult height changes with the number of growth measurements of height and bone age. A marked increase in precision is obtained when a single height measurement is added to a single bone age measurement. A small improvement with increasing numbers of height or bone age measures occurs thereafter. This is apparent also from Table 3, where there are low correlations between the higher order coefficients and adult height. Also, the addition of bone age gives a greater improvement than the addition of an extra height measurement. Where there are no height measurements extra bone age measurements contribute nothing since only the bone age intercept term is correlated with adult height. Note that the standard error depends only on the values of the explanatory variables in the basic model, ie age, and not the actual measurements.

Discussion

We have shown how adult height can be predicted efficiently using any set of height or bone age measurements made during growth. This contrasts with previous procedures (eg Tanner *et al.*, 1982) which predict from one or two growth measurements at fixed ages, and are thus less flexible. A procedure analogous to the present one is given by Bock (1986). This uses non-linear growth curves, however, and has not been extended to handle covariates or multiple response measures. Such curves are also rather inflexible in that, while being able to graduate height growth reasonably well, they are generally unsuitable for measurements such as weight or skinfold which have no well-defined upper asymptote.

By fitting a model which includes further adult measurements we can predict these also. We can also add further growth period measurements to the model to improve the accuracy of prediction, although beyond a certain point these would not be expected to add a great deal.

A major shortcoming in the present analyses is the relatively small sample size which limits both the number of random parameters which can be fitted and the accuracy of their estimates. If several data sets are combined in an analysis, then allowance should be made for overall differences in the fixed and possibly in the random parts. Thus, the mean adult height will differ, in general, between groups, and the between subject variance may also differ. Likewise, various growth period parameters may differ. This will be the case not only for different countries but also for different social groups, family sizes, pubertal stages, menarche occurrence etc. It may also be necessary to recognize the existence of a secular trend between data collection and the time of use by adding an adjustment to the mean adult height.

There are a number of technical issues also, which remain to be studied. These include the robustness of the computational procedures in the case of non-Guassian distributions and the modelling of measurements made close together in time. In the present analysis we have used yearly measurements, so that the assumption, in equation (2), that the level-1 random terms are independent, can be expected to hold. This will not be true, however, when measurements are made close together, and some form of dependency structure will be needed (see Goldstein, 1987).

Finally, it should be emphasized that all predictions are subject to error, and the presentation of a confidence interval about a prediction is as important as the single value prediction itself.

References

Bock, R. D. (1986): Unusual growth patterns in the Fels data. In *Human growth: a multidisciplinary review*, ed A. Demirjian. London and Philadelphia: Taylor and Francis.

Goldstein, H. (1986): Efficient statistical modelling of longitudinal data. *Ann. Hum. Biol.* **13**, 129–41.

Goldstein, H. (1987): *Multilevel models in educational and social research*. London, Griffin: New York, Oxford University Press.

Strenio, J., Weisberg, H. I. & Bryk, A. S. (1983): Empirical Bayes estimation of individual growth curve parameters and their relationship to covariates. *Biometrics* **39**, 71–86.

Tanner, J. M., Whitehouse, R. H., Cameron, N., Marshall, W. A., Healy, M. J. R. & Goldstein, H. (1982): Assessement of skeletal maturity and prediction of adult height (TW2 method). London: Academic Press.

APPENDIX

The full model can be written as:

$$y_{ij} = (1 - \omega_{ij})\delta_{ij} \sum_{t=0}^{1} \alpha_{tij} x^{t}_{ij} + (1 - \omega_{ij})(1 - \delta_{ij}) \sum_{t=0}^{3} \beta_{tij} x^{t}_{ij} + \omega_{ij}\gamma_{i} + \omega_{ij}\lambda z_{ij}$$

The coefficients are composed of a fixed (mean) term and random terms as follows, where u represents level-2 variation and e level-1 variation.

$\alpha_{0ij} = \alpha_0 + u_{10i} + e_{10ij}$
$\alpha_{1ij} = \alpha_1 + u_{11i}$
$\beta_{0ij} = \beta_0 + u_{20i} + e_{20ij}$
$\beta_{1ij} = \beta_1 + u_{21i}$
$\beta_{2ij} = \beta_2 + u_{22i}$
$\beta_{3ij} = \beta_3 + u_{23i}$
$\gamma_i = \gamma + u_3$
$\omega_{ij} = 1$ if adult height, 0 otherwise
$\delta_{ij} = 1$ if bone age measurement, 0 otherwise

The first term in the model is for bone age, the second for height, the third for adult height and the fourth is for the subgroup using the (0,1) dummy variable z.

Auxology 88. Perspectives in the science of growth and development ed. J. M. Tanner.
© Smith-Gordon, 1989.