ELSEVIER SCIENCE
IRELAND

International Journal of Bio-Medical Computing
35 (1994) 39–46

International Journal of
Bio-Medical
Computing

PC program for growth prediction in the two-stage polynomial growth curve model

Ingrid (Ying-Yueh) Y. Guo[a], Emet D. Schneiderman*[b], Charles J. Kowalski[c], Stephen M. Willis[b]

[a]*Department of Public Health Sciences,* [b]*Department of Oral and Maxillofacial Surgery and Pharmacology, Baylor College of Dentistry, 3302 Gaston Avenue, Dallas, TX 75246, USA*
[c]*Department of Biologic and Materials Science, The University of Michigan, Ann Arbor, MI 48109, USA*

(Received 9 March 1993; accepted 9 April 1993)

Abstract

We consider the problem of growth prediction in the context of the two-stage (or random coefficients) one-sample polynomial growth curve model and provide a PC program, written in GAUSS386i, to perform the associated computations. The problem considered is that of estimating the value of the measurement under consideration for a 'new' individual at the *T*th time point given measurements on that individual at $T-1$ previous points in time and the values of the measurement on N 'similar' individuals at all T time points. The times of measurement $t_1, t_2, \ldots, t_T$ need not be equally spaced, but we assume that each of the N individuals comprising the normative sample were measured at these times. The method and the program are illustrated using the data set previously considered (Schneiderman and Kowalski, *Am J Phys Anthrop,* 67 (1985) 323–333) consisting of mandibular ramus height measurements (in mm) for 12 male rhesus monkeys at $T=5$ yearly intervals (coded 1, 2, 3, 4, and 5). Results are compared with those obtained under a less restrictive set of assumptions concerning the covariance matrix of the observations than is made in the context of the two-stage model. It is seen that the accuracies of prediction of the two methods, for this and other data sets, are quite close, suggesting that the less restrictive model may be preferred in many situations.

Key words: Longitudinal data; Polynomial growth curves; Prediction; PC program

* Corresponding author.

SSDI 0020-7101(93)00840-E

1. Introduction

We have previously described the two-stage polynomial growth curve model [1], and documented a number of the advantages which accrue when orthogonal (or orthonormal) polynomials are used to define the within-individual (time) design matrix in this and other longitudinal data-analytic contexts [2]. While the program described in this paper allows several forms of the time design matrix, due to the advantages mentioned above and to allow easy comparison with Ware and Wu [3], who developed the theory behind our approach, we use the notation and formulae appropriate for orthonormal polynomials (so $\Phi'\Phi = I$ where Φ is the time design matrix). Within this framework, the structure and distributional assumptions of the two-stage model may be summarized by

$$\begin{aligned} \boldsymbol{x}_i|\alpha_i &\sim \text{MVN}\,(\Phi\alpha_i,\ \sigma^2\mathbf{I}) \\ \alpha_i &\sim \text{MVN}\,(\alpha,\ \Lambda) \\ \boldsymbol{x}_i &\sim \text{MVN}\,(\Phi\alpha,\ \Phi\Lambda\Phi' + \sigma^2\mathbf{I}) \end{aligned}$$

where $\boldsymbol{x}_i|\alpha_i$ denotes the conditional distribution of $\boldsymbol{x}_i$ given α_i and, for example, $\alpha_i \sim \text{MVN}\,(\alpha, \Lambda)$ is read 'α_i has a multivariate normal distribution with mean or expected value α and covariance matrix Λ.' Our earlier papers may be consulted for detailed descriptions of these quantities and of a PC program which:

(a) determines the lowest degree polynomial adequate to provide an acceptable fit to this model;
(b) estimates the parameters α_i, α, σ^2 and Λ;
(c) computes confidence intervals for the elements of α;
(d) provides confidence bands for the average growth curve (AGC);
(e) produces plots of the AGC and its associated confidence bands.

The purpose of the present paper is to extend this methodology — and our program — to accomodate growth prediction, i.e. to allow the user to estimate the value of the measurement under consideration for a 'new' individual at the Tth time point given measurements on that individual at $T - 1$ previous points in time and given the values of the measurements on N 'similar' individuals at all T time points. The times of measurement $t_1, t_2, \ldots, t_T$ need not be equally spaced, but we assume that the time design matrix, Φ, is the same for each of the $N + 1$ individuals (the N individuals comprising the normative sample and the individual whose growth we wish to predict), i.e. that the $t_1, t_2, \ldots, t_T$ are not individual-specific.

Formally, we may state the problem as follows: given

$$X_{N\times T} = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1T} \\ x_{21} & x_{22} & \cdots & x_{1T} \\ \vdots & \vdots & \vdots & \vdots \\ x_{N1} & x_{N2} & \cdots & x_{NT} \end{bmatrix} \tag{1}$$

where x_{ij} is the value of the measurement for the *i*th individual ($i = 1, 2, \ldots, N$) at time t_j ($j = 1, 2, \ldots, T$), and given the first $T - 1$ entries of

$$\boldsymbol{x}_\nu = \begin{bmatrix} x_{\nu 1} \\ x_{\nu 2} \\ \cdots\cdot \\ x_{\nu, T-1} \\ x_{\nu T} \end{bmatrix} \tag{2}$$

estimate the value of $\boldsymbol{x}_{\nu T}$. Our exposition is based on Ref. 3.

2. Prediction of $x_{\nu T}$

The general solution of the prediction problem was previously outlined in Ref. 4 where it was shown how $x_{\nu T}$ could be predicted in the context of Rao's [5] one-sample polynomial growth curve model. The solution in terms of the two-stage model (which is also due to Rao, but will be referred to here as the two-stage model to distinguish it from his earlier model) will be briefly sketched as follows. We partition the vector $\boldsymbol{x}_\nu$ into its known and unknown parts, namely,

$$\boldsymbol{x}_\nu = \begin{bmatrix} x_{\nu 1} \\ \vdots \\ x_{T-1} \\ -- \\ x_{\nu T} \end{bmatrix} = \begin{bmatrix} \boldsymbol{x}_\nu^* \\ -- \\ x_{\nu T} \end{bmatrix} \tag{3}$$

so that $\boldsymbol{x}_\nu^*$ is $(T - 1) \times 1$, the observed values for the ν^{th} individual, and $x_{\nu T}$ is the (scalar) quantity to be predicted. The time design matrix $\boldsymbol{\Phi}$ is partitioned similarly into the $(T - 1) \times P$ matrix $\boldsymbol{\Phi}_1$ and the $1 \times P$ matrix $\boldsymbol{\Phi}_2$, namely,

$$\boldsymbol{\Phi} = \begin{bmatrix} 1 & \phi_1(t_1) & \cdots\cdot & \phi_D(t_1) \\ \vdots & \vdots & \cdots\cdot & \vdots \\ 1 & \phi_1(t_{T-1}) & \cdots\cdot & \phi_D(t_{T-1}) \\ -- & -- & -- & -- \\ 1 & \phi_1(t_T) & \cdots\cdot & \phi_D(t_T) \end{bmatrix} = \begin{bmatrix} \boldsymbol{\Phi}_1 \\ -- \\ \boldsymbol{\Phi}_2 \end{bmatrix} \tag{4}$$

where D is the degree of the final polynomial growth curve model [1] and $P = D + 1$ the number of parameters in this model.

In terms of these submatrices the covariance matrix of the $\boldsymbol{x}_i$, $\Sigma = \boldsymbol{\Phi}\Lambda\boldsymbol{\Phi}' + \sigma^2 I$ can then be written

$$\Sigma = \begin{bmatrix} \boldsymbol{\Phi}_1\Lambda\boldsymbol{\Phi}_1' + \sigma^2 I_{T-1} & | & \boldsymbol{\Phi}_1\Lambda\boldsymbol{\Phi}_2' \\ -- & & -- \; -- \\ \boldsymbol{\Phi}_2\Lambda\boldsymbol{\Phi}_1' & | & \boldsymbol{\Phi}_2\Lambda\boldsymbol{\Phi}_2' + \sigma^2 \end{bmatrix} \tag{5}$$

and from standard multivariate normal theory (e.g. Ref. 6, p. 442) the conditional mean and variance of $x_{\nu T}$ given $\mathbf{x}_\nu{}^*$ are

$$E(\mathbf{x}_{\nu T}\,|\,x_\nu{}^*) = \Phi_2\alpha + (\Phi_2\Lambda\Phi_1')(\Phi_1\Lambda\Phi_1' + \sigma^2 I)^{-1}(\mathbf{x}_\nu{}^* - \Phi_1\alpha) \qquad (6)$$

and

$$V(x_{\nu T}\,|\,\mathbf{x}_\nu{}^*) = \sigma^2 + \Phi_2\Lambda\Phi_2' - (\Phi_2\Lambda\Phi_1')\,(\Phi_1\Lambda\Phi_1' + \sigma^2 I)^{-1}\,(\Phi_1\Lambda\Phi_2') \qquad (7)$$

An estimator of $x_{\nu T}$ is then obtained by substituting estimates [1] of α, σ^2, and Λ in Eq. 6; the estimated prediction variance results when these substitutions are made in Eq. 7. An approximate 95% confidence interval for $x_{\nu T}$ is

$$\hat{E}(x_{\nu T}\,|\,\mathbf{x}_\nu{}^*) \pm 2\sqrt{\hat{V}(x_{\nu T}\,|\,\mathbf{x}_\nu{}^*)}$$

3. The program

The program is essentially a combination of two previously documented programs: one which estimates the parameters in the two-stage model [1] and one which computes the predicted values and their variances [4]. Since this documentation is available and since the program is interactive, only a brief overview of the program operation is provided here.

The user is prompted for the name and location of an ASCII (or GAUSS) data set of the form of Eq. 1, containing the observations for the N individuals at the T times of measurement. She is then requested to enter the values for the 'new' individual at the first $T - 1$ time points. The output includes D, the smallest degree adequate to fit the data; the estimated values of the elements of α and their corresponding 95% confidence intervals; the 95% confidence bands for the AGC at each time of measurement; the estimated value of $x_{\nu T}$; and an approximate 95% confidence interval for this quantity. The AGC and its confidence bands are then plotted and the predicted value for the first 'new' individual is highlighted. The user is then asked whether or not another prediction is to be made. If yes, the user is prompted for the observed values of the second 'new' individual at the first $T - 1$ time points. The numerical output at this stage consists only of the predicted value and the prediction interval. The graphical output is the same as above. The program continues in this fashion until the user responds in the negative ('N') to the question concerning another individual's prediction.

Finally, as an option, the user may choose to apply the leave-one-out method to her data set. This method is described in the following section (see also Ref. 4).

4. An example

Our example is based on the data set previously considered in Ref. 7, consisting of mandibular ramus height measurements (in mm) for 12 male rhesus monkeys at $T = 5$-yearly intervals (coded 1, 2, 3, 4, and 5). This data set was also used in Ref.

4 to illustrate prediction in the context of Rao's [5] polynomial growth curve model. Use of these data thus allows the comparison of the two approaches. We employ the leave-one-out (LOO) method in which $N = 12$ predictions are made; we leave one monkey out of the computations involving the normative sample at each stage and predict his value at $T = 5$. Since the actual values at $T = 5$ are known for each monkey, a comparison of these values with the predicted values provides some insight into the accuracy with which predictions are being made. This method was used in growth prediction contexts by Rao [8,9] and other applications were indicated by Lachenbruch [10]. The results for the predictions based on Rao's model [4] and the two-stage model are shown in Table 1.

The root mean square error (RMSE) of prediction [4] for Rao's model is 0.56, while for the two-stage model, RMSE = 0.68. It is seen that there is little difference between the two methods for this data set. In this case, however, the degrees of the polynomials adequate to fit the two models differed ($D = 2$ for Rao; $D = 3$ for two-stage), and the initial estimate of Λ was not positive definite, requiring a correction [1], which may call into question the appropriateness of the two-stage model in this situation [11]. Accordingly, we present the results for another data set, one that has been extensively studied in the context of growth prediction by Rao [8,9], and one for which the same degree polynomial is adequate for both models ($D = 1$) and the estimate of Λ in the two-stage model is positive definite. It consists of ramus heights of $N = 20$ boys measured at ages 8, 8.5, 9 and 9.5 years. We predict the values at 9.5 years of age given the earlier measurements. The results are shown in Table 2. For Rao's method, RMSE = 0.65; for two-stage, RMSE = 0.72. We see again that there is little to choose between the two methods and, in fact, Rao's method is slightly better than the two-stage model even though there is no reason to suspect the appropriateness of the two-stage model for these data.

Table 1
Results for the predictions on the ramus height measurements of 12 rhesus monkeys, based on Rao's model and the two-stage model

Monkey	$T = 5$ Actual	Predicted (Rao)	Predicted (two-stage)
1	35.8	35.6	36.2
2	43.5	43.4	42.9
3	38.9	39.4	39.5
4	44.4	43.5	43.8
5	37.9	38.6	38.8
6	43.8	44.0	43.4
7	43.1	43.2	43.4
8	44.0	44.8	44.8
9	43.8	44.0	44.3
10	42.0	42.1	42.1
11	43.8	42.9	42.3
12	43.8	44.4	44.2

Table 2
Results for the predictions on the ramus height measurements of 20 boys, aged 8–9.5 years, based on Rao's model and the two-stage model

Actual value at 9.5 years	Predicted (Rao)	Predicted (two-stage)
49.7	49.8	49.9
48.4	48.6	48.6
48.5	48.8	48.6
47.2	47.1	46.7
49.3	49.7	49.8
53.7	53.9	54.0
54.5	55.2	55.6
52.7	51.1	50.9
54.4	53.3	54.0
48.3	48.1	48.7
51.9	52.6	52.4
55.5	54.6	54.1
55.0	54.5	54.5
49.8	50.1	50.1
51.8	52.1	52.1
53.3	53.6	53.7
49.5	49.3	49.1
55.3	55.8	56.1
48.4	49.0	49.2
51.8	52.9	52.6

Rao [8,9] studied the performance of seven different predictors on this data set. He obtained RMSE's ranging from 0.70 to 0.80, so that both methods considered above are competitive with his, at least in so far as this data set is concerned. We have also compared the methods on a number of other data sets. The general conclusion is that all methods produce generally comparable results — Rao's method is (slightly) better in some cases, two-stage in others. Since Rao's method makes fewer assumptions than the two-stage model (specifically, for Rao, the covariance matrix, Σ, of the observations, x, is arbitrary; while in the two-stage model it has the special structure $\Sigma = \Phi\Lambda\Phi' + \sigma^2 I$, it may be preferred for general use. For more details concerning the structure of the two-stage model, see Ref. 1.

5. Discussion

Here we consider the results for the monkey data set in more detail: in particular, how they may be related to the phenomenon of tracking [3]. It is seen that the predictions are quite close for this data set, both for the approach based on Rao's model [4] and the two-stage model [1]. This occurs despite the fact that these monkeys do not track especially well as judged by the values of the tracking indices we have implemented, these being an index based on the kappa statistic [12], and two forms of the index developed by Foulkes and Davis [13], denoted here by FDI [14] and FDII

[15]. In fact, their estimated values and the 95% confidence intervals for the corresponding parameters are:

Kappa (with three tracks):	0.24242 ± 0.12990
FDI:	0.39394 ± 0.14902
FDII (with $D = 2$):	0.53030 ± 0.13018

This is somewhat unexpected. Indeed, Ware and Wu [3] essentially equate tracking with the ability to predict. Obtaining accurate predictions even when other indices indicate a lack of tracking is perhaps a reflection of the facts that tracking indices measure particular aspects of growth patterns, and small values do not preclude prediction. One can expect that prediction will be quite good when tracking is in evidence, but tracking is not a necessary condition for the ability to predict. For a more detailed discussion, see Ref. 16.

6. Acknowledgement

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

7. Appendix: Computer implementation

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 progams require a 80386- or 80486-based personal computer (PC) running the MS-DOS operating system (version 5.0 or higher is recommended, although versions as low as 3.3 will suffice). 80386 computers must also be equipped with a 80387 math coprocessor. At least 4 Mb of memory is required, and must be available to GAUSS386i, i.e. not in use by memory resident programs such as Windows. EGA or VGA graphic capabilities are required to display the color graphics; VGA or SVGA is suggested to optimally display 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.

8. References

1 Ten Have TR, Kowalski CJ and Schneiderman ED: 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.

2 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.

3 Ware JH and Wu MC: Tracking: prediction of future values from serial measurements, *Biometrics*, 37 (1981) 427–437.

4 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.

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

6 Rao CR: *Linear Statistical Inference and Its Applications*, Wiley, New York, 1965.

7 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.

8 Rao CR: Prediction of future observations with special reference to linear models. In *Multivariate Analysis IV* (Ed: PR Krishnaiah), North-Holland, Amsterdam, 1977, pp. 193–208.

9 Rao CR: Prediction of future observations in growth curve models, *Stat Sci*, 2 (1987) 434–471.

10 Lachenbruch PA: *Discriminant Analysis*, Hafner, New York, 1975.

11 Carter AL and Yang MCK: Large-sample inference in random-coefficient regression models, *Comm Stat — Theory Methods*, 8 (1986) 2507–2526.

12 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.

13 Foulkes MA and Davis CE: An index of tracking for longitudinal data, *Biometrics*, 37 (1981) 439–446.

14 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.

15 Schneiderman ED, Willis SM, Kowalski CJ and Ten Have TR: A GAUSS program for computing the Foulkes-Davis tracking index for polynomial growth curves, *Int J Biomed Comput*, 32 (1993) 35–43.

16 Kowalski CJ and Schneiderman ED: Tracking: concepts, methods and tools, *Hum Evol*, in press.