Comput. Biol. Med. Vol. 23, No. 2, pp. 149–154, 1993
Printed in Great Britain

0010-4825/93 $6.00+.00
© 1993 Pergamon Press Ltd

LONGER-TERM GROWTH PREDICTION USING GAUSS

Emet D. Schneiderman,[*] Stephen M. Willis,[*] Charles J. Kowalski[†] and Thomas R. Ten Have[‡]

[*] Department of Oral and Maxillofacial Surgery, Baylor College of Dentistry, 3302 Gaston Avenue, Dallas, TX 75246, U.S.A.; and Departments of [†] Oral Biology, and [‡] Biostatistics, The University of Michigan, Ann Arbor, MI 48109, U.S.A.

(*Received* 9 *June* 1992; *in revised form* 27 *October* 1992; *received for publication* 23 *November* 1992)

Abstract—In several areas of biomedicine, one needs to predict future measurements for a growing individual on the basis of longitudinal data. Here we consider the problem of estimating the values of a given measurement for a particular individual at $T-T^*$ points in time, given T^* observations on that individual, and all T values for a sample of N "similar" individuals. This extends our previous discussion [Schneiderman *et al.*, *Comput. Biol. Med.* **22**, 181–188 (1992)], which was limited to the case $T^*=T-1$, to longer-term predictions. We again make a user-friendly GAUSS program available to perform the associated computations. Examples illustrating the use of the program and the accuracy of the predictions it provides are included.

Longitudinal studies Growth Prediction Polynomials PC program

INTRODUCTION

In an earlier paper [1] we described a method for a simple form of growth prediction in the context of Rao's [2] one-sample polynomial growth curve model and provided a PC program, written in GAUSS, to perform the associated computations. We considered the situation in which longitudinal data on N individuals were available at T time points (not necessarily equally spaced) and we wished to predict the value of a "new" individual at time T given observations on that individual at the first $T-1$ time points. The purpose of the present paper is to extend that discussion, and our program, to longer-term prediction, i.e. we suppose that the new individual has been measured at a subset (T^*) of the total (T) time points, $1 \leq T^* \leq T-1$, and it is desired to predict that subject's unknown values at the remaining $R=T-T^*$ points in time.

THE MODEL

For a description of Rao's model, see [3]; for its use in the simple form of growth prediction described above, see [1]. Here we consider the situation in which we have a $N \times T$ data matrix X consisting of the values of the measurement under consideration for each of N individuals from a specified population at time $t_1, t_2, \ldots, t_T$. Using this information and the values of the measurement for a new individual *from that population* at the first T^* of these times, we wish to predict the remaining $R=T-T^*$ values for this new individual. We denote by $\mathbf{x}_v$ the $T \times 1$ vector of observations for individual v. As in our earlier publication we begin by partitioning $\mathbf{x}_v$ into its known and unknown parts, namely,

$$\mathbf{x}_v = \left[\begin{array}{c} x_{v1} \\ \vdots \\ x_{vT^*} \\ \hline x_{v,T^*+1} \\ \vdots \\ x_{vT} \end{array}\right] = \left[\begin{array}{c} \mathbf{x}_v^* \\ \hline \mathbf{x}_{vR} \end{array}\right] \tag{1}$$

so that $\mathbf{x}_v^*$ is $T^* \times 1$, the observed values for the vth individual, and $\mathbf{x}_{vR}$ is the $R \times 1$ vector of values to be predicted. The time-design and sample covariance matrices, W $(T \times P)$ and S $(T \times T)$, are partitioned similarly [1]:

$$W = \begin{bmatrix} 1 & t_1 & \cdots & t_1^D \\ 1 & t_2 & \cdots & t_2^D \\ \vdots & \vdots & & \vdots \\ 1 & t_T & & t_T^D \end{bmatrix} = \left[\frac{W_1}{W_2}\right] \tag{2}$$

and

$$S = \left[\frac{S_{11} \mid S_{12}}{S_{21} \mid S_{22}}\right] \tag{3}$$

The sample covariance matrix S is based on the vector of sample means. In (2), W_1 is $T^* \times P$ and W_2 is $R \times P$. In (3), S_{11} is $T^* \times T^*$, $S_{12} = S_{21}'$ is $T^* \times R$ and S_{22} is $R \times R$. P is the number of parameters needed to adequately fit a polynomial to the average growth curve (AGC). If this polynomial is of degree D, $P = D + 1$. Having determined D [3], the P coefficients of the polynomial describing the AGC are estimated by

$$\hat{\tau} = (W'S^{-1}W)^{-1}W'S^{-1}\bar{\mathbf{x}}, \tag{4}$$

where $\bar{\mathbf{x}}$ is the $T \times 1$ vector of means at each time point.

Then given X and $\mathbf{x}_v^*$, we estimate (predict) the remaining values by

$$\hat{\mathbf{x}}_{vR} = E(\mathbf{x}_{vR} \mid \mathbf{x}_v^*) = W_2\hat{\tau} + S_{21}S_{11}^{-1}(\mathbf{x}_v^* - W_1\hat{\tau}) \tag{5}$$

and the estimated prediction variance is

$$\hat{V}(\mathbf{x}_{vR} \mid \mathbf{x}_v^*) = S_{22} - S_{21}S_{11}^{-1}S_{12}. \tag{6}$$

Equations (5) and (6) represent, respectively, the conditional mean and variance of $\mathbf{x}_{vR}$ given $\mathbf{x}_v^*$. $\hat{V}(\mathbf{x}_{vR} \mid \mathbf{x}_v^*)$ is an $R \times R$ matrix with the prediction variances of the R predicted values on the diagonals. These can be used to construct confidence intervals for the predicted values as indicated in [1].

We turn now to a brief description of the program and illustrate its use using the data considered previously in [3].

THE PROGRAM

This program, LTPRED, is entirely similar to the one described in detail in [1]; the major differences are the dimensions of the corresponding vectors and matrices to accommodate our desire to predict more than a single value. The user is first "prompted" (provided with a menu to respond to) for information concerning the data matrix X and the times of measurement. Then the user is asked to enter the values of $\mathbf{x}_v^*$. A period (.) signals that T^* has been reached, i.e. that all the available observations for the current individual have been entered. The corresponding predicted values, their variances, and approximate 95% prediction intervals are printed. The growth profile of the individual, along with highlighted predicted values are plotted against the backdrop of the estimated AGC and its 95% confidence band. The user is then asked whether to not another individual's values are to be predicted. The program continues in this manner until the response to this question is negative at which time the user is given the opportunity to save the expanded data matrix consisting of X augmented by one or more rows containing the values of the $\mathbf{x}_v$ [both the observed and predicted values as in (1)] for each of the "new" individuals considered. The above description of the program should suffice to document its use. For a more detailed discussion, see [1]. We turn now to several examples selected to give some indication of the accuracy of the predictions made using this method.

SOME EXAMPLES

Consider the data set used in [3] consisting of the values of mandibular ramus height in each of $N = 12$ young male rhesus monkeys at $T = 5$ (equally spaced) time points which, for convenience, is reproduced below and included with copies of the program which can be obtained as indicated in the Appendix.

Monkey	Time 1	Time 2	Time 3	Time 4	Time 5
1	25.2	29.0	33.6	35.2	35.8
2	27.3	32.1	37.0	41.8	43.5
3	26.3	30.7	36.1	38.0	38.9
4	26.0	34.5	39.0	42.3	44.4
5	25.5	29.5	34.4	38.3	37.9
6	28.2	32.5	36.3	42.3	43.8
7	25.4	33.4	38.0	42.7	43.1
8	27.2	34.8	37.2	44.0	44.0
9	26.0	34.5	38.0	43.5	43.8
10	28.5	33.8	38.0	39.2	42.0
11	27.0	31.2	36.0	41.7	43.8
12	26.0	33.0	40.2	42.5	43.8

We consider three examples based on these data. In the first, consider a (hypothetical) monkey with some subset of the (rounded) mean values for the group at each time point, namely, 26.6, 32.4, 37.0, 41.0, 42.1. Using our program to predict the remaining values given the indicated subset of measurements, we obtained the results summarized below:

Mean monkey	26.6	32.4	37.0	41.0	42.1
Given 26.6 (at t_1)	...	32.94	37.65	40.74	42.21
Given 26.6, 32.4	...	...	37.20	40.13	41.59
Given 26.6, 32.4, 37.0	...	...	...	40.11	41.48
Given 26.6, 32.4, 37.0, 41.0	...	...	...	...	42.23

It is seen that all the predictions are quite accurate, even those based on a single observation. This is perhaps not too surprising given that the "mean monkey's" growth pattern is so well-behaved. We next consider another new (hypothetical) monkey with each of the measurements at the first four time points being somewhat below the mean. Suppose this new monkey presents with some subsets of the observations 26, 30, 35, 40 at the first four times of measurement. The results obtained from our program are summarized below:

New monkey	26	30	35	40	t_5?
Given 26 (at t_1)	...	32.60	37.49	40.37	41.51
Given 26, 30	...	...	35.37	37.48	38.60
Given 26, 30, 35	...	...	...	37.43	38.39
Given 26, 30, 35, 40	...	...	...	...	40.55

In successive rows above we are given increasing information and asked to predict fewer remaining values. Note that all predictions remain below the mean value for the $N = 12$ monkeys.

We next consider one of the actual monkeys in the data set, namely, monkey #1 with (actual) measurements 25.2, 29.0, 33.6, 35.2, 35.8 at the five times of measurement. Following the methodology outlined above for the remaining $N = 11$ monkeys, the predicted values given various subsets of these measurements are indicated below:

Monkey #1	25.2	29.0	33.6	35.2	35.8
Given 25.2 (at t_1)	...	32.08	37.27	39.87	40.57
Given 25.2, 29.0	...	...	34.74	36.41	37.09
Given 25.2, 29.0, 33.6	...	...	...	36.26	36.45
Given 25.2, 29.0, 33.6, 35.2	...	...	...	...	35.56

The strategy of omiting a monkey in order to assess the accuracy of prediction for that monkey is known as the leave-one-out (LOO) method. Rao [4, 5] has used this method in the context of growth prediction. See [6] for a good general description and other applications. Below we summarize the results obtained when the LOO strategy is applied to each of the 12 monkeys in turn when predicting the final two observations from the first three. The actual values are shown in parentheses:

Monkey	t_4	t_5
1	36.26 (35.2)	36.45 (35.8)
2	39.78 (41.8)	41.70 (43.5)
3	38.27 (38.0)	39.66 (38.9)
4	42.53 (42.3)	43.65 (44.4)
5	36.86 (38.3)	37.42 (37.9)
6	40.06 (42.3)	42.12 (43.8)
7	41.30 (42.7)	41.99 (43.1)
8	42.55 (44.0)	43.57 (44.0)
9	42.39 (43.5)	43.09 (43.8)
10	41.59 (39.2)	44.13 (42.0)
11	38.74 (41.7)	40.36 (43.8)
12	41.16 (42.5)	43.31 (43.8)

It is seen that, for each monkey, the predicted values are quite close to the actual. A measure of the accuracy of prediction is the root mean square error, RMSE [1]. The value of this quantity for t_4 is RMSE = 1.683 while for t_5 it is 1.483. Similar results were obtained for different numbers, R, of remaining values to be predicted.

We can also provide some indication of the prediction *variance* associated with the use of this method on this data set. As can be seen from equation (6), the prediction variance depends only on the dimension of $\mathbf{x}_v^*$, not the particular values in $\mathbf{x}_v^*$, i.e. these are the same for each monkey. We show below the estimated prediction variances in our example for various choices of $R(T^*)$, the number of remaining observations to be predicted (the number of observations):

R	T^*	t_2	t_3	t_4	t_5
4	1	3.40	3.36	6.64	6.67
3	2	...	1.06	2.35	2.32
2	3	...	...	2.33	1.99
1	4	...	...	...	0.34

Thus, for example, when $R = 2$ $(T^* = 3)$, the predicted values at times 4 and 5 have prediction variances of 2.33 and 1.99, respectively. One would expect, in general, that the variances in each column will decrease: the fewer values that need to be predicted, the smaller the variances associated with the prediction. Stated otherwise, if you want to predict the value at t_5, it is best to have all of the preceeding four observations. Similarly, the values in each row will generally increase: for a fixed number of observations, the further into the future you wish to predict, the greater the uncertainty of the prediction. These expectations will, of course, not *always* be realized in practice, e.g. predicting what will happen near puberty may well be more difficult than predicting what will happen at a more distant point in time when growth patterns have "settled down".

DISCUSSION

We acknowledge that some users will realize that this program can be used to "fill-in" longitudinal data sets that are incomplete (contain missing data) due to drop-outs. Several approaches have been proposed for analyzing data sets with incomplete repeated measurements [7–11]; however, the practical application of these awaits the development of appropriate software. We have, however, developed a GAUSS program explicitly for estimating randomly occurring missing values within longitudinal data sets [12] that builds on the approach presented here and in [1]. Unlike the present approach, geared towards prediction, this other method incorporates a procedure [13] for restoring the noise into the estimated observations that is removed by the smoothing functions. While not essential in the context of prediction, this adjustment is important when estimating values for subsequent analysis (i.e. filling in missing observations), otherwise standard errors of the estimates will be systematically underestimated. When imputing more than a handful of missing values, the use of the program outlined in [12] is perhaps preferable. Another approach to dealing with missing data using the EM algorithm has also been implemented as a GAUSS program [14]. Additionally, it must be emphasized that that whether one estimates the values of the missing observations or uses an analysis which can accommodate missing data, it is important to be sure that the drop-outs have occurred "at random", i.e. that the incomplete measurement sequences are not atypical.

Diggle [11] gives a good discussion and outlines a test which may be used to check on this assumption. In any case, we suggest that users prudently limit the numbers of observations filled-in using the methods of this paper.

Having made these caveats, we do allow users to save the enlarged data set which results when several individuals' observations have been estimated. That is, if n longitudinal sequences have been estimated, the original $N \times T$ data matrix, X, can be augmented to produce an $(N+n) \times T$ matrix which can then be read into any of our (or others') programs requiring complete data. The user is prompted as to his/her decision to save this augmented matrix. We suggest that the total number of items to be estimated (nR) should be small relative to the total number of actual observations (NT), certainly $<10\%$; and that the effect(s) of using the enlarged data set can be at least partially assessed by comparing the outputs from a given program obtained when X and the augmented matrix are used in turn. We would also suggest that the user report the proportion of observations that have been estimated whenever the augmented data set is subjected to analysis. Finally, it is recommended that one only predicts future values on the basis of the *complete original observations*, i.e. it is not advisable to use vectors with estimated values to estimate values of other new subjects.

SUMMARY

In the context of several medical and dental specialties, as well as in growth studies, one may want to estimate future measurements for a growing individual on the basis of longitudinal data. An investigator or practitioner may also wish to predict the long-term response of a subject to a treatment for which he/she has normative longitudinal data in hand. In this paper we consider a method for estimating the values of a given measurement for a particular individual at $T-T^*$ points in time given T^* observations on that individual and all T values for a sample of N "similar" individuals. This extends our previous discussion [1] which was limited to the case $T^*=T+1$, to longer-term predictions. We again made a user-friendly GAUSS program available to perform the associated computations. Examples illustrating the use of the program and the accuracy of the predictions it provides were included. It was seen that quite accurate predictions were possible, even with small sample sizes, and even when given relatively few observations, T^*.

It was noted that this methodology could be used to fill-in longitudinal data sets with missing data due to drop-outs, but that care needs to be taken to ensure that the incomplete measurement sequences are not atypical.

Acknowledgement—This research was supported by grant DE08730 from the National Institute of Dental Research.

REFERENCES

1. E. D. Schneiderman, S. M. Willis, C. J. Kowalski and T. R. Ten Have, A PC program for growth prediction in the context of Rao's polynomial growth curve model, *Comput. Biol. Med.* **22**, 181–188 (1992).
2. C. R. Rao, Some problems involving linear hypotheses in multivariate analysis, *Biometrika* **46**, 49 (1959).
3. E. D. Schneiderman and C. J. Kowalski, Implementation of Rao's one-sample polynomial growth curve model using SAS, *Am. J. Phys. Anthrop.* **67**, 323 (1985).
4. C. R. Rao, Prediction of future observations with special reference to linear models, *Multivariate Analysis IV*, P. R. Krishnaiah, Ed., pp. 193–208. North-Holland, Amsterdam (1977).
5. C. R. Rao, Prediction of future observations in growth curve models, *Statist.* **2**, 434 (1987).
6. P. A. Lachenbruch, *Discriminant Analysis.* Hafner, New York (1975).
7. H. Crepeau, J. Koziol, N. Reid and Y. S. Yuh, Analysis of incomplete multivariate data from repeated measurement experiments, *Biometrics* **41**, 505 (1985).
8. R. H. Jones, Serial correlation in unbalanced mixed models, *Bull. Int. Stat. Inst.* **52**, 105 (1987).
9. M. G. Kenward, A method for comparing profiles of repeated measurements, *Appl. Statist.* **36**, 296 (1987).
10. G. O. Zerbe and S. H. Walker, A randomization test for comparison of groups of growth curves with different polynomial design matrices, *Biometrics* **33**, 653 (1977).
11. P. J. Diggle, Testing for random dropouts in repeated measurement data, *Biometrics* **45**, 1255 (1989).
12. E. D. Schneiderman, C. J. Kowalski and S. M. Willis, Regression imputation of missing values in longitudinal data sets, *Int. J. Biomed. Comput.* **32**, 121 (1993).
13. R. J. A. Little and D. B. Rubin, The analysis of social science data with missing values, *Modern Methods of Data Analysis*, J. Fox and J. S. Long, Eds, pp. 374–409. Sage, Newbury Park, CA (1990).

14. R. S. Schoenberg, MISS: a computer program for the estimation of moments and imputation of missing data when observations are incomplete, RJS Software, P.O. Box 2883, Kensington MD 20895 (1988).

About the Author—EMET DAN SCHNEIDERMAN received the B.A. and M.A. in Anthropology from Northwestern University in 1978, and a Ph.D. in Biological Anthropology from The University of Michigan in 1985. While at The University of Michigan he was affiliated with the Center for Human Growth and Development and began conducting research in the area of craniofacial growth. In collaboration with Joseph Mudar, Schneiderman developed an integrated software system for the analysis of cephalometric radiographs (X-rays of the head). While on the orthodontics faculty of the University of Detroit School of Dentistry from 1985 to 1988, Dr Schneiderman created a computerized cephalometry laboratory. In 1988 Dr Schneiderman went to the Baylor College of Dentistry in Dallas where he is the director of research for the department of oral and maxillofacial surgery and pharmacological sciences. Dr Schneiderman and co-investigator Dr Charles Kowalski have been funded by NIH/NIDR from 1988 to 1993 to conduct the biostatistical research from which this paper issued. Dr Schneiderman has more than 45 publications including two chapters and the monograph, *Facial Growth in the Rhesus Monkey*, published by Princeton University Press in 1992.

About the Author—STEPHEN M. WILLIS received the B.S. degree in Mathematics from the University of Texas at Arlington in 1987. Mr Willis has over 15 years of experience in clinical toxicology and is currently operations manager of a regional toxicology laboratory in Dallas. Mr Willis is also the lead programmer/systems analyst for the NIH/NIDR grant on longitudinal statistical methods with Drs Kowalski and Schneiderman. Mr Willis has played a major role in the development of user-friendly interfaces for programs that have broad applications in the biomedical sciences.

About the Author—CHARLES J. KOWALSKI received the B.S. in Mathematics from Roosevelt University in Chicago in 1962, the M.S. in Statistics from Michigan State University in 1965, and a Ph.D. in Biostatistics from The University of Michigan in 1968. Dr Kowalski then joined the faculty of the Department of Oral Biology at the University of Michigan School of Dentistry. Dr Kowalski served as the assistant director of the university's Statistical Research Laboratory from 1971 to 1978 and research scientist at the Dental Research Institute from 1978 to the present, and directed the institute's biometrics laboratory. He has been full professor of dentistry and statistician at The University of Michigan since 1978. At various times Dr Kowalski has served as a consultant to the National Football League, Park, Davis and Co., Nymegen University, Lanchester Cleft Palate Clinic, the Department of Antiquities of the University of Alexandria in Egypt, the U.S. Veterans Administration and the Eastman Dental Center. Dr Kowalski has published more than 180 scientific papers, including numerous chapters and the book *A Mixed-longitudinal Interdisciplinary Study of Growth and Development*, published by Academic Press in 1979. Dr Kowalski's research has focused on the application of statistical methods to dental and oral research with special emphasis on measurement processes, their validity, reliability and calibration. Longitudinal data analysis and the computer implementation of polynomial growth curve models have also been and continue to be a major thrust of his research. Drs Kowalski and Schneiderman have been funded by NIH/NIDR from 1988 to 1993 to study and implement biostatistical methods for the analysis of longitudinal data in the form of user-friendly microcomputer programs.

About the Author—THOMAS R. TEN HAVE received the B.A. in Statistics in 1981, an M.P.H. in Biostatistics in 1982, and Ph.D. in Biostatistics in 1991, all from The University of Michigan. Dr Ten Have served as a biostatistician at the Center for Human Growth and Development and the Statistical Research Laboratory of The University of Michigan from 1985 to 1991. Dr Ten Have is now at the Center for Biostatistics and Epidemiology at the Hershey (Pennsylvania) Medical Center. In addition to the analysis of longitudinal craniofacial growth and development data, Dr Ten Have also conducts research on the longitudinal analysis of categorical data. Dr Ten Have has co-authored more than 25 scientific papers.

APPENDIX. COMPUTER IMPLEMENTATION

A full set of PC programs for longitudinal data analysis, including this program, can be obtained on 5.25" or 3.5" diskettes (please request type) by sending $25 to defray the cost of handling and licensing fees. These programs require a 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 display optimally the graphic results. Runtime modules are supplied with the programs so that no additional software (i.e. compiler or interpreter) is required to run these programs. One can create and edit ASCII data sets for use by these programs using the full screen editor supplied with MS-DOS version 5.0. The programs are written and compiled using GAUSS386i, version 3.0, require no additional installation or modification, and are run with a single command. When requesting the programs, address inquiries to the corresponding author and make checks payable to Baylor College of Dentistry.