Int J Biomed Comput, 32 (1993) 121–133
Elsevier Scientific Publishers Ireland Ltd.

REGRESSION IMPUTATION OF MISSING VALUES IN LONGITUDINAL DATA SETS[a]

EMET D. SCHNEIDERMAN[a], CHARLES J. KOWALSKI[b] and STEPHEN M. WILLIS[a]

[a]*Department of Oral and Maxillofacial Surgery, Baylor College of Dentistry, Gaston Ave, Dallas, TX,* [b]*Center for Statistical Consultation and Research, The University of Michigan (USA)*

(Received August 20th, 1992)
(Accepted September 25th, 1992)

A stand-alone, menu-driven PC program, written in GAUSS, which can be used to estimate missing observations in longitudinal data sets is described and made available to interested readers. The program is limited to the situation in which we have complete data on N cases at each of the planned times of measurement t_1, t_2,..., t_T; and we wish to use this information, together with the non-missing values for n additional cases, to estimate the missing values for those cases. The augmented data matrix may be saved in an ASCII file and subsequently imported into programs requiring complete data. The use of the program is illustrated. Ten percent of the observations in a data set consisting of mandibular ramus height measurements for $N = 12$ young male rhesus monkeys measured at $T = 5$ time points are randomly discarded. The augmented data matrix is used to determine the lowest degree polynomial adequate to fit the average growth curve (AGC); the regression coefficients are estimated and confidence intervals for them are determined; and confidence bands for the AGC are constructed. The results are compared with those obtained when the original complete data set is used.

Key words: Regression; Longitudinal studies; Missing data; PC program

Introduction

We have written and made available a number of PC programs which perform various longitudinal data analyses [1–9]. An overview of these is given in Ref. 10 and details concerning the tracking programs are provided in Ref. 11. Each of these programs accepts unequally spaced time points t_1, t_2,..., t_T, but none, at the moment, allows missing data. That is, the data matrix consisting of the values of the measurement under consideration for N individuals at T time points, viz.,

$$\mathbf{X} = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1T} \\ x_{21} & x_{22} & \cdots & x_{2T} \\ \vdots & \vdots & \cdots & \vdots \\ x_{N1} & x_{N2} & \cdots & x_{NT} \end{bmatrix} \tag{1}$$

Correspondence to: Emet D. Schneiderman, Department of Oral and Maxillofacial Surgery, Baylor College of Dentistry, 3302 Gaston Ave, Dallas, TX 75246, USA.

0020-7101/93/$06.00 © 1993 Elsevier Scientific Publishers Ireland Ltd.
Printed and Published in Ireland

can have no missing entries. The purpose of the present paper is to describe, illustrate and make available a user-friendly, menu-driven, PC program that can be used to estimate missing entries in Eqn. (1), thus allowing the inclusion of individuals with one or more missing data points. We assume that the study is planned so that observations are to be made at the fixed set of times t_1, t_2,...,t_T but that n individuals were not observed at one or more of these times. Our objective is to estimate these values.

The discussion (and program) is limited to the case where we have complete data on N cases and wish to use this information, and the observed values for a given one of n cases with missing data, to estimate the missing values for that case. Thus the program is a direct extension of our programs for prediction [12,13]: Here we allow the missing values to occur anywhere in the vector of observations for an individual, not just at the end. We again use the conditional expectation of the missing values given the observed values for that individual and for the N cases with complete data, to fill-in the missing data points. To fix notation, we consider that the total data set is structured as

$$\underset{(N+n)\times T}{\mathbf{X}} = \begin{bmatrix} X_1 \\ X_2 \end{bmatrix} \tag{2}$$

where $\mathbf{X}_1$ is $N \times T$, consisting of N cases with complete data; $\mathbf{X}_2$ is $n \times T$, consisting of n cases, each of which contains one or more missing data points. Examples will follow.

Other approaches to missing data problems exist. A general review of the literature up to 1966 is given in Ref. 14. A simple taxonomy of missing data problems is given in Ref. 15. More recent general discussions are also available [16–18]. Missing data problems due to non-response in sample surveys are described in Refs. 19 and 20. A comparison of the kinds of problems which arise in cross-sectional and longitudinal studies is made in Ref. 21.

In the context of longitudinal data analysis, some of the approaches which have been used include:

(i) Delete cases with missing data
(ii) Use specialized, noniterative techniques which allow missing data
(iii) Use the EM algorithm
(iv) Fit individual curves; use the estimated regression coefficients as the basic data
(v) Imputation methods

The first of these is self-explanatory and is probably the 'safest' way to proceed [22], provided that the data are missing at random and that the proportion of cases with missing data is small. There are situations, however, when the majority of cases have one or more missing data points and discarding these would result in a sample size so small as to preclude any analysis. Discarding those subjects with incomplete data is easy to carry out and may be satisfactory in situations where only a small number of subjects present with missing data. This approach can, however, lead to serious biases [17] and is not very efficient in the sense that all the information present in partially recorded observation vectors is ignored.

Examples of (ii) are included in Refs. 23–26. These are promising approaches, but their routine application awaits implementation. Some require that one make (typically strong) assumptions concerning the structure of the correlations between the repeated measurements. Others are large-sample procedures whose small-sample operating characteristics are yet to be studied in detail. For example, Ref. 26 extends the two-stage polynomial growth curve model [4] to allow missing data when N or T is large. Programs implementing these methods are under development and it will be of considerable interest to compare their results with those obtained by the method considered here, as well as the results from iterative methods based on the EM algorithm.

The EM (estimation/maximization) algorithm is a general method for estimating missing data in a variety of situations. The EM algorithm was given its name by Dempster et al. [27] who presented the general theory for the algorithm and a number of examples. A GAUSS program is already available [28]. Applications to growth curve problems are considered by Laird et al. [29].

Examples of (iv) include those reported by Dawson et al. and Zerbe [30,31]. Here individual growth curves are fit and a τ_i thus obtained for each case. The τ_i are then used as the basic data and one may, e.g., construct confidence bands for the average growth curve (AGC) and/or the individual growth curves [30] and derive standards for growth, growth velocity and acceleration on this basis [31]. Different groups of individuals may also be compared using this method [32]. We note in passing that the data used in Ref. 30 consisted of a sample of $n = 11$ achondroplasic children, none of which were measured at all $T = 13$ time points ($t = 0,1,..., 12$ months). Our program cannot be used in such situations. Rather, it is intended for use when n is small relative to N, and when the pattern of missing data may reasonably be described as 'incidental.'

In (v), missing data values are filled-in and the resultant completed data set is analyzed by standard methods. Common forms of imputation include hot deck imputation, where actual observations from other subjects are substituted for missing values; mean imputation, where mean values computed from the complete data are substituted; and regression imputation, where the missing values for a subject are estimated either by predicted values from the conditional regression as described above, or by predicted values from the regression on the known values from that subject alone. This latter form of regression imputation is equivalent to the approach taken by Dawson et al. and Zerbe [30,31] and is not considered further here. An important consideration when any imputation technique is used is how to modify subsequent analyses to allow for the differing status of the real and imputed values. Our approach to this question is outlined in the following section.

Methods

We assume that each row, $\mathbf{x}'$, of $\mathbf{X}$ has a multivariate normal distribution with mean or expected value

$$E(\mathbf{x}) = \mathbf{W}\tau \qquad (3)$$

and (arbitrary) covariance matrix Σ. In (2), $\mathbf{W}$ is the $P \times T$ within-individual (or time) design matrix used to fit a polynomial of degree $D = P - 1$ to the data and τ is the $P \times 1$ vector of polynomial regression coefficients for AGC, viz.,

$$\mathbf{W} = \begin{bmatrix} 1 & t_1 & \dots & t_1^D \\ 1 & t_2 & \dots & t_2^D \\ \vdots & \vdots & \dots & \vdots \\ 1 & t_T & \dots & t_T^D \end{bmatrix} \quad \text{and} \quad \tau = \begin{bmatrix} \tau_1 \\ \tau_2 \\ \vdots \\ \tau_P \end{bmatrix} \tag{4}$$

The smallest degree polynomial adequate to fit the data for the N cases is determined as outlined in [1,13]. We now consider one of the n cases with missing data. If m of the entries of $\mathbf{x}$ are missing, we write the model in partioned form as

$$\mathbf{x} = \begin{bmatrix} \mathbf{x}_1 \\ \mathbf{x}_2 \end{bmatrix} = \begin{bmatrix} \mathbf{W}_1 \\ \mathbf{W}_2 \end{bmatrix} \tau \tag{5}$$

where $\mathbf{x}_1$ is $(T - m) \times 1$, $\mathbf{x}_2$ is $m \times 1$, $\mathbf{W}_1$ is $(T - m) \times P$ and $\mathbf{W}_2$ is $m \times P$. In (5), the entries of $\mathbf{x}$ are rearranged (if necessary) so that $\mathbf{x}_1$, contains the values actually observed and $\mathbf{x}_2$ the missing data points. We also partition the $T \times T$ sample covariance matrix, S, as in (5), viz.,

$$\mathbf{S} = \begin{bmatrix} \mathbf{S}_{11} & \mathbf{S}_{12} \\ \mathbf{S}_{21} & \mathbf{S}_{22} \end{bmatrix} \tag{6}$$

so that $\mathbf{S}_{11}$ is $(T - m) \times (T - m)$, $\mathbf{S}_{12} = \mathbf{S}'_{21}$ is $(T - m) \times m$ and $\mathbf{S}_{22}$ is $m \times m$. It should be noted that S is computed using the N cases with complete data, but it is partitioned in accordance with the pattern of missing data for the case under consideration, i.e., $\mathbf{S}_{11}$ contains the covariances of the measurements actually observed for that case; $\mathbf{S}_{12}$ (and $\mathbf{S}'_{21} = \mathbf{S}_{12}$) the covariances between the observed and missing observations; and $\mathbf{S}_{22}$ the covariances among the missing observations. S is computed just once, but it is rearranged and partitioned as many times as there are distinct patterns of missing data.

Having determined D, the P coefficients of τ are estimated by

$$\hat{\tau} = (\mathbf{W}'\mathbf{S}^{-1}\mathbf{W})^{-1}\, \mathbf{W}'\mathbf{S}^{-1}\bar{\mathbf{x}} \tag{7}$$

where $\bar{\mathbf{x}}$ is the $T \times 1$ vector of means at each time point [13]. We then estimate $\mathbf{x}_2$ by

$$\hat{\mathbf{x}}_2 = \mathbf{W}_2\hat{\tau} + \mathbf{S}_{21}\mathbf{S}_{11}^{-1}\,(\mathbf{x}_1 - \mathbf{W}_1\hat{\tau}) \tag{8}$$

which is of the same form as used in Refs. 12 and 13. When imputing missing data, however, this should be modified. It is important to realize that since we intend to use estimated values as observations, some adjustment needs to be made. For reasons given in [18] and considered in more detail in the Discussion, we take

$$\tilde{\mathbf{x}}_2 = \hat{\mathbf{x}}_2 + \mathbf{e} \tag{9}$$

where **e** is the residual from a randomly selected complete case. That is, focus for the moment on the N complete cases. We have $\hat{\tau}$ computed from these cases and the fitted values for a given one of these cases is computed by

$$\hat{\mathbf{x}} = \mathbf{W}\hat{\tau} \tag{10}$$

with the corresponding residual

$$\mathbf{e} = \mathbf{x} - \hat{\mathbf{x}} \tag{11}$$

The elements of this randomly selected residual corresponding to the missing elements in the case whose values are being estimated are added to $\mathbf{x}_2$. A detailed example is considered below. We first give a brief description of the program.

The Program

The menu-driven program is invoked by issuing the command

gsruni missing

The program menu appears and the user is prompted for the location of the data file, which can be in a different directory than the program itself. The indicated directory is searched and the names of those data files with the extension '.ASC' are displayed. It is assumed that these files have the following properties:

(i) They are rectangular — rows corresponding to subjects, columns to data values;
(ii) They are in ASCII format with one or more spaces separating the data values;
(iii) Missing values are coded by periods (".");
(iv) Subjects with missing values can be in any row (not just at the end);
(v) Missing data are missing at random; and
(vi) $N > T$ and $N > n$.

The user then highlights the file of choice using the cursor arrow keys and selects the file with the return key. He/she then supplies various simple pieces of information concerning the structure of the data and the manner in which it is to be analyzed. These options, implemented in the form of questions, include:

(i) Are the observations made at equally spaced time points? If the time points are equally spaced, the user is given the option of starting at 1 (default) or any other starting time. The interval between time points is entered next. If the time points are not equally spaced, the user is prompted for each of the time points. Fractional (e.g., 4.75) and negative (e.g., for centering the data) values are allowed; and

(ii) What level of significance (e.g., 0.05) is to be used in determining the smallest degree, D, adequate to fit the AGC?

Some Examples

A simple example based on the data in Schneiderman and Kowalski [1], where polynomials of degree $D = 2$ were fit to mandibular ramus height measurements (mm) for $N = 12$ young male rhesus monkeys at times coded $t = 1, 2, 3, 4, 5$ will clarify some of the above notions. For convenience, these data are reproduced below

$$\mathbf{X} = \begin{bmatrix} 25.2 & 29.0 & 33.6 & 35.2 & 35.8 \\ 27.3 & 32.1 & 37.0 & 41.8 & 43.5 \\ 26.3 & 30.7 & 36.1 & 38.0 & 38.9 \\ 26.0 & 34.5 & 39.0 & 42.3 & 44.4 \\ 25.5 & 29.5 & 34.4 & 38.3 & 37.9 \\ 28.2 & 32.5 & 36.3 & 42.3 & 43.8 \\ 25.4 & 33.4 & 38.0 & 42.7 & 43.1 \\ 27.2 & 34.8 & 37.2 & 44.0 & 44.0 \\ 26.0 & 34.5 & 38.0 & 43.5 & 43.8 \\ 28.5 & 33.8 & 38.0 & 39.2 & 42.0 \\ 27.0 & 31.2 & 36.0 & 41.7 & 43.8 \\ 26.0 & 33.0 & 40.2 & 42.5 & 43.8 \end{bmatrix}$$

For these data we found

$$\bar{\mathbf{X}}' = [\ 26.6, \quad 32.4, \quad 37.0, \quad 41.0, \quad 42.1\ ]$$

$$\mathbf{S} = \begin{bmatrix} 1.1756 & 0.7236 & 0.3145 & 0.7314 & 1.3718 \\ 0.7236 & 3.8451 & 2.9876 & 4.2699 & 4.6906 \\ 0.3145 & 2.9876 & 3.4451 & 3.4747 & 4.1221 \\ 0.7314 & 4.2699 & 3.4747 & 7.0954 & 7.2139 \\ 1.3718 & 4.6906 & 4.1221 & 7.2139 & 8.2715 \end{bmatrix}$$

and a $D = 2$ polynomial was adequate to fit the AGC (p = 0.1356).

When $T = 5$ and $D = 2$ ($P = 3$), the time design matrix is

$$W = \begin{bmatrix} 1 & 1 & 1 \\ 1 & 2 & 4 \\ 1 & 3 & 9 \\ 1 & 4 & 16 \\ 1 & 5 & 25 \end{bmatrix}$$

and so the estimated regression coefficients for the AGC are

$$\hat{\tau} = (\mathbf{W}'\mathbf{S}^{-1}\mathbf{W})^{-1}\,\mathbf{W}'\mathbf{S}^{-1}\bar{\mathbf{x}} = \begin{bmatrix} 18.5572 \\ 8.8189 \\ -0.8198 \end{bmatrix}$$

The user may wish to center the time-design matrix, **W**. In this example, the centered time points would be -2, -1, 0, 1, 2 and we would have

$$\mathbf{W} = \begin{bmatrix} 1 & -2 & 4 \\ 1 & -1 & 1 \\ 1 & 0 & 0 \\ 1 & 1 & 1 \\ 1 & 2 & 4 \end{bmatrix}$$

Centering does not effect the predicted values, but does reduce the multicollinearity inherent (especially for large values of D) in the uncentered form of **W** [6]. One is therefore encouraged to select the option to center the time points (in this case the program would assign equally spaced time points starting at -2 and incrementing by 1).

For purposes of illustration, we use the uncentered form of **W** and suppose that the measurements of the 11th monkey are missing at times $t = 2$ and $t = 4$ and that the 12th monkey is missing at times $t = 1$, $t = 4$ and $t = 5$, i.e., that the observation vectors for these monkeys are:

$$\mathbf{x}_{11} = \begin{bmatrix} 27.0 \\ ? \\ 36.0 \\ ? \\ 43.8 \end{bmatrix} \quad \text{and} \quad \mathbf{x}_{12} = \begin{bmatrix} ? \\ 33.0 \\ 40.2 \\ ? \\ ? \end{bmatrix}$$

The input data file with $N = 10$ and $n = 2$ would then be prepared as shown below with periods (".") representing the missing data points.

```
25.2   29.0   33.6   35.2   35.8
27.3   32.1   37.0   41.8   43.5
26.3   30.7   36.1   38.0   38.9
26.0   34.5   39.0   42.3   44.4
25.5   29.5   34.4   38.3   37.9
28.2   32.5   36.3   42.3   43.8
25.4   33.4   38.0   42.7   43.1
27.2   34.8   37.2   44.0   44.0
26.0   34.5   38.0   43.5   43.8
28.5   33.8   38.0   39.2   42.0
27.0     .    36.0     .    43.8
  .    33.0   40.2     .      .
```

The monkeys with missing data are then considered in turn: the **W** and **S** matrices are rearranged and partitioned to reflect the patterns of missing data for each. To facilitate comparison, we use the **S** computed above; it would, of course, change slightly if cases 11 and 12 were omitted from the computation. We also use $\hat{\mathbf{r}}$ as computed previously.

For monkey 11, we have

$$\mathbf{W}_1 = \begin{bmatrix} 1 & 1 & 1 \\ 1 & 3 & 9 \\ 1 & 5 & 25 \end{bmatrix} \quad \mathbf{W}_2 = \begin{bmatrix} 1 & 2 & 4 \\ 1 & 4 & 16 \end{bmatrix} \quad \text{and} \quad \mathbf{x}_1 = \begin{bmatrix} 27.0 \\ 36.0 \\ 43.8 \end{bmatrix}$$

and

$$\mathbf{S}_{11} = \begin{bmatrix} 1.1756 & 0.3145 & 1.3718 \\ 0.3145 & 3.4451 & 4.1221 \\ 1.3718 & 4.1221 & 8.2715 \end{bmatrix}$$

$$\mathbf{S}_{12} = \begin{bmatrix} 0.7236 & 0.7314 \\ 2.9876 & 3.4747 \\ 4.6906 & 7.2139 \end{bmatrix} = \mathbf{S}'_{21}$$

and

$$\mathbf{S}_{22} = \begin{bmatrix} 3.8451 & 4.2699 \\ 4.2699 & 7.0954 \end{bmatrix}$$

Then $\mathbf{x}_2$ is estimated by

$$\hat{\mathbf{x}}_2 = \mathbf{W}_2\hat{\mathbf{r}} + \mathbf{S}_{21}\mathbf{S}_{11}^{-1}(\mathbf{x}_1 - \mathbf{W}_1\hat{\mathbf{r}}) = \begin{bmatrix} 32.633679 \\ 42.642340 \end{bmatrix}$$

To illustrate the computation of $\hat{\mathbf{x}}_2$, we randomly select one of the $N = 10$ cases with complete data. If, e.g. this were no. 7 with $\mathbf{x}'_7 = [25.4,33.4,38.0,42.7,43.1]$ then the fitted values for no. 7 are

$$\hat{\mathbf{x}}_7 = \mathbf{W}\hat{\mathbf{r}} = \begin{bmatrix} 26.5563 \\ 32.9158 \\ 37.6357 \\ 40.7160 \\ 42.1567 \end{bmatrix}$$

Then

$$\mathbf{e} = \mathbf{x}_7 - \hat{\mathbf{x}}_7 = \begin{bmatrix} -1.1563 \\ 0.4842 \\ 0.3643 \\ 1.9840 \\ 0.9433 \end{bmatrix}$$

From this we choose the 2nd and 4th elements so that

$$\tilde{\mathbf{x}}_2 = \hat{\mathbf{x}}_2 + \mathbf{e} = \begin{bmatrix} 32.633679 \\ 42.642340 \end{bmatrix} + \begin{bmatrix} 0.4842 \\ 1.9840 \end{bmatrix} = \begin{bmatrix} 33.12 \\ 44.63 \end{bmatrix}$$

Thus the final values for case 11 are

$$\begin{bmatrix} 27.0 \\ 33.1 \\ 36.0 \\ 44.6 \\ 43.8 \end{bmatrix}$$

The procedure for case 12 is the same. In this case

$$\mathbf{W}_1 = \begin{bmatrix} 1 & 2 & 4 \\ 1 & 3 & 9 \end{bmatrix}, \quad \mathbf{W}_2 = \begin{bmatrix} 1 & 1 & 1 \\ 1 & 4 & 16 \\ 1 & 5 & 25 \end{bmatrix}, \quad \text{and} \quad \mathbf{x}_1 = \begin{bmatrix} 33.0 \\ 40.2 \end{bmatrix}$$

and

$$\mathbf{S}_{11} = \begin{bmatrix} 3.8451 & 2.9876 \\ 2.9876 & 3.4451 \end{bmatrix}$$

$$\mathbf{S}_{12} = \begin{bmatrix} 0.7236 & 4.2699 & 4.6906 \\ 0.3145 & 3.4747 & 4.1221 \end{bmatrix} = \mathbf{S}_{21}$$

and

$$\mathbf{S}_{22} = \begin{bmatrix} 1.1756 & 0.7314 & 1.3718 \\ 0.7314 & 7.0954 & 7.2139 \\ 1.3718 & 7.2139 & 8.2715 \end{bmatrix}$$

We compute $\hat{\mathbf{x}}_2$ and $\tilde{\mathbf{x}}_2 = \hat{\mathbf{x}}_2 + \mathbf{e}$, where $\mathbf{e}$ is the residual for another randomly selected case, as before.

The complete, filled-in 12 × 5 data matrix $\mathbf{X}$ can now be named and saved in an

ASCII file. On the screen, **X**'s real values appear in a color different from that of the imputed values.

The example considered above was intended to illustrate the partitioning of **x**, **W** and **S**; and how $\hat{\mathbf{x}}_2$ is modified to produce the final value $\tilde{\mathbf{x}}_2$. It was convenient to use $\hat{\tau}$ and **S** as computed in Ref. 1 for this purpose. The following example will show how the technique works in practice. We use the same data set, but randomly delete 10% (6 data points) of the observations, viz., we delete x_{15}, x_{33}, x_{61}, x_{64}, x_{93}, and $x_{12,5}$. We have $N = 7$, $n = 5$ and the data are entered as

27.3	32.1	37.0	41.8	43.5
26.0	34.5	39.0	42.3	44.4
25.5	29.5	34.4	38.3	37.9
25.4	33.4	38.0	42.7	43.1
27.2	34.8	37.2	44.0	44.0
28.5	33.8	38.0	39.2	42.0
27.0	31.2	36.0	41.7	43.8
25.2	29.0	33.6	35.2	.
26.3	30.7	.	38.0	38.9
.	32.5	36.3	.	43.8
26.0	34.5	.	43.5	43.8
26.0	33.0	40.2	42.5	.

The data set is filled-in using the estimated values $\hat{x}_{15} = 33.9$, $\hat{x}_{33} = 35.8$, $\hat{x}_{61} = 27.3$, $\hat{x}_{64} = 42.6$ and $\hat{x}_{12,5} = 47.1$

From the data set with imputed values, we find

$$\bar{\mathbf{x}}' = [26.477,\ 32.417,\ 36.964,\ 40.983,\ 42.182]$$

$$\mathbf{S} = \begin{bmatrix} 0.9761 & 0.7170 & 0.3708 & 0.6466 & 1.3098 \\ 0.7170 & 3.8452 & 3.0494 & 4.2721 & 5.4448 \\ 0.3708 & 3.0494 & 3.5166 & 3.5564 & 5.7512 \\ 0.6466 & 4.2721 & 3.5564 & 7.1740 & 8.6933 \\ 1.3098 & 5.4448 & 5.7512 & 8.6933 & 12.6964 \end{bmatrix}$$

and

$$\hat{\tau}' = [18.68,\ 8.617,\ -0.758]$$

It is seen that these values are in generally good agreement with those computed using the complete data set, shown earlier.

We can also compare the confidence intervals for the elements of τ and the confidence bands for the AGC. The confidence intervals are

Coefficient	Complete	Augmented
τ_1	(16.41, 20.70)	(16.94, 20.41)
τ_2	(7.211, 10.43)	(7.491, 9.742)
τ_3	(-1.055, -0.5571)	(-0.9589, -0.5571)

and the confidence bands for the AGC at the five time points are

Time	Complete	Augmented
1	(24.90, 28.21)	(25.12, 27.94)
2	(31.45, 34.38)	(32.15, 33.60)
3	(35.39, 39.88)	(36.31, 39.10)
4	(37.84, 43.59)	(38.51, 43.51)
5	(38.60, 45.72)	(38.75, 46.87)

It is seen that the confidence intervals for the elements of τ are somewhat narrower when the missing data points are imputed, as are the confidence bands for the AGC (with the exception of $t = 5$), but the differences are relatively minor.

Discussion

While the estimates of the elements of τ are quite similar in the example under consideration, the confidence intervals and bands are somewhat narrower when missing data are imputed than when the original, complete data set is used. The extent to which this may be a general phenomenon will require further investigation. It does seem clear that the direct use of $\hat{\mathbf{x}}_2$ as computed in Eqn. (8) would tend to oversmooth the data. The addition of random residuals is an attempt to restore some of the noise to the system, noise which was smoothed out by the regression function. However, it is less clear that this restores the appropriate amount of noise. We suggest that the problem will not be serious as long as the use of the program is limited to situations in which n is small relative to N. See Ref. 18 for a more detailed discussion.

It is also important that the missing data be 'missing at random,' i.e., that if **x** has some missing entries, this fact does not depend on the values of the elements which were actually observed. A good discussion is given by Rubin [33] and a procedure which can be used to test for this when data are missing due to droputs is given by Diggle [34].

Acknowledgement

Supported by grant DE08730 from the National Institute of Dental Research.

Appendix — Computer Implementation

This program can be obtained on a 5.25″ or 3.5″ diskette (please request type) by sending $25 to defray the cost of handling and licensing fees. The progam requires 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 megabytes of memory is required, and must be available to GAUSS386, 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 program so that no additional software (i.e., compiler or interpreter) is required to run this program. One can create and edit ASCII data sets for use by this program using the full screen editor supplied with MS-DOS version 5.0. The program is written in GAUSS386, version 3.0, requires no additional installation or modification and is run with a single command. When requesting the program, address inquiries to the corresponding author and make checks payable to Baylor College of Dentistry.

References

1 Schneiderman ED and Kowalski CJ: Implementation of Rao's one-sample polynomial growth curve model using SAS, *Am J Phys Anthrop*, 67 (1985) 323–333.

2 Schneiderman ED and Kowalski CJ: Implementation of Hills' growth curve analysis for unequal-time intervals using GAUSS, *Am J Hum Biol*, 1 (1989) 31–42.

3 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.

4 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.

5 Schneiderman ED, Willis SM, Ten Have TR and Kowalski CJ: Rao's polynomial growth curve model for unequal-time intervals: A menu-driven GAUSS program, *Int J Biomed Comput*, 29 (1991) 235–244.

6 Ten Have TR, Kowalski CJ and Schneiderman ED: PC program for obtaining orthogonal polynomial regression coefficients for use in longitudinal data analysis, *Am J Hum Biol*, 4 (1992) 403–416.

7 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.

8 Schneiderman ED, Willis SM, Kowalski CJ and Ten Have TR: PC program for comparing tracking indices in several independent groups, *Am J Hum Biol*, 4 (1992) 399–401.

9 Ten Have TR, Kowalski CJ, Schneiderman ED and Willis SM: Two programs for performing multigroup longitudinal analyses, *Am J Phys Anthrop*, 88 (1992) 251–254.

10 Kowalski CJ: Data analysis in craniofacial biology with special emphasis on longitudinal studies. *Clef Palate-Craniofacial J*, in press.

11 Kowalski CJ and Schneiderman ED: Tracking: Concepts, methods and tools. In: *Multivariate Methods in Bioanthropology* (Eds: GN van Vark, W Schaafsma and RS Corruccini), *Human Evol.* (Special Issue) in press.

12 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.

13 Schneiderman ED, Willis SM, Kowalski CJ and Ten Have TR: Longer-term growth prediction using GAUSS, *Comput Biol Med*, in press.

14 Afifi AA and Elashoff RM: Missing observations in multivariate statistics I. Review of the literature, *J Am Stat Assn*, 61 (1966) 595–604.

15 Hartley HO and Hocking RR: The analysis of incomplete data, *Biometrics*, 27 (1971) 783–823.

16 Little RJA and Rubin DB: Incomplete data. In: *Encyclopedia of Statistical Sciences*, vol. 4. (Eds: S Kotz, NL Johnson and CB Read), Wiley, New York, 1983.

17 Little RJA and Rubin DB: *Statistical Analysis with Missing Data*, Wiley, New York, 1987.

18 Little RJA and Rubin DB: The analysis of social science data with missing values. In: *Modern Methods of Data Analysis*, (Eds: J Fox and JS Long), Sage, Newbury Park, CA, 1990.

19 Little RJA: Models for nonresponse in sample surveys, *J Am Stat Assn*, 77 (1982) 237–250.

20 Sande IG: Imputation in surveys: coping with reality, *Am Stat*, 36 (1982) 145–152.

21 Rovine MJ and Delaney M: Missing data estimation in developmental research. In: *Statistical Methods in Longitudinal Research*, vol. I, (Eds: A von Eye), Academic Press, New York, 1990.

22 Afifi AA and Clark V: *Computer-Aided Multivariate Analysis*, Lifetime Learning, Belmont, CA, 1984.

23 Crepeau H, Koziol J, Reid N and Yuh YS: Analysis of incomplete multivariate data from repeated measurement experiments. *Biometrics*, 41 (1985) 505–514.

24 Jones RH: Serial correlation in unbalanced mixed models. *Bull Int Stat Inst*, 52 (1987) 105–122.

25 Kenward MG: A method for comparing profiles of repeated measurements. *Appl Stat*, 36 (1987) 296–308.

26 Carter RL and Yang MCK: Large-sample inference in random-coefficient regression models. *Comm Stat*, 8 (1986) 2507–2562.

27 Dempster AP, Laird NM and Rubin DB: Maximum likelihood from incomplete data via the EM algorithm. *J Roy Stat Soc*, B39 (1977) 1–38.

28 Schoenberg RS: MISS: *A Computer Program for the Estimation of Moments and Imputation of Missing Data when Observations are Incomplete*, RJS Software, 26250 196th Place SE, Kent WA 98042.

29 Laird N, Lange N and Stram D: Maximum likelihood computations with repeated measures: Application of the EM algorithm. *J Am Stat Assoc*, 82 (1987) 97–105.

30 Dawson DV, Todorov AB and Elston RC: Confidence bands for the growth of head circumference in achondroplastic children during the first year of life. *Am J Med Genet*, 7 (1980) 529–536.

31 Zerbe GO: A new nonparametric technique for constructing percentiles and normal ranges for growth curves determined from longitudinal data. *Growth*, 43 (1979) 263–272.

32 Zerbe GO and Walker SH: A randomization test for comparison of groups of growth curves with different polynomial design matrices. *Biometrics*, 33 (1977) 653–657.

33 Rubin DB: Inference and missing data. *Biometrika*, 63 (1976) 581–582.

34 Diggle PJ: Testing for random dropouts in repeated measurement data. *Biometrics*, 45 (1989) 1255–1258.