Estimation of Missing Values in Longitudinal Data Sets
Using Regression Methods in Biological Research

G. Tamer KAYAALP and Fatma ÇEVÝK
\textsuperscript{1}Faculty of Agriculture, \textsuperscript{2}Faculty of Fisheries, Çukurova University, Turkey

Abstract: In biological researches long period data sets were called longitudinal data. But statistical analysis cannot be applied when one or more observations are missing. For the estimation of missing values in longitudinal data sets, regression methods were used. The records from 30 water temperature values (3 station \times 10 water temperature at different months) were taken. At the end estimated values for incomplete observations were similar to the model for the full completed observations. This results showed that the regression method for the incomplete observations can be used in similar cases by researchers.

Key words: Longitudinal data, regression method, missing value, biological research

Introduction
For the case of two or independent variables the efficiency of regression method depends upon the correlation between the independent variables and the proportion of missing observations.

Kayaalp (1999), has illustrated a method for multiple regression models with missing data. An experiment is planned so that observations are to be made at the fixed set of times \( t_1, t_2, \ldots, t_n \) but that \( n \) individuals were not observed at one or more times. The objective is to estimate these values. The discussion is limited to the case where complete data is available 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. The statistical literature on missing data does not answer this question in general. In most articles data is ignored after being is assumed accidental in one sense or another. In some articles such as those concerned with the multivariate normal (Anderson, 1957; Affifi and Elashoff, 1968; Hocking and Smith, 1988; Hartley and Hocking, 1971).

Weighted least squares analyses described by Grizzle et al. (1959). They have been developed the analysis of incomplete longitudinal categorial data. Haltovsky (1988) and Hartley and Hocking (1971), considered several processes that can cause missing values. This notion is more complicated than it seems and it was discussed more by Rubin (1976). Although the properties of these procedures have been extensively compared in simulation studies (Timm, 1970; Beale and Little, 1976; Donner and Rosner, 1982).

Materials and Methods
30 water temperature values (3 station \times 10 water temperature at different months) were taken into account from three station at Seyhan River in Adana - Turkey. It was assumed that each row, \( X_i \), of \( X \) has a multi variate normal distribution with mean or expected value,

\[
E(X) = W \gamma
\]

And \( \gamma \) arbitrary covariance matrix, \( X \) is the \( P \times T \) within-individual (or time) design matrix used to fit a polynomial of degree \( P = T - 1 \) to data and \( y \) is the \( P \times 1 \) vector of polynomial regression coefficients.

\[
W = \begin{bmatrix}
1 & 1 & \ldots & 1 \\
1 & t_1 & \ldots & t_1 \gamma \\
\vdots & \vdots & \ddots & \vdots \\
1 & t_n & \ldots & t_n \gamma
\end{bmatrix}
\]

We now consider one of the \( n \) cases with missing data. If \( m \) of the entries of \( X \) are missing, we write the model \( n \) partitioned form as

\[
X = \begin{bmatrix}
X_1 \\
X_2
\end{bmatrix} = \begin{bmatrix}
W_1 \\
W_2
\end{bmatrix} \gamma
\]

Where \( X_1 \) is \( (T - m) \times m \times 1 \), \( W_1 \) is \( (T - m) \times P \) and \( W_2 \) is \( m \times P \). In equation 3 the entries of \( X \) are rearranged (if necessary) so that \( X_1 \), contains the values actually observed and \( X_2 \) the missing data points. We also partition the \( T \times T \) sample covariance matrix, \( S \), as in equation 4.

\[
S = \begin{bmatrix}
S_{11} & S_{12} \\
S_{21} & S_{22}
\end{bmatrix}
\]

So that \( S_{11} \) is \( (T - m) \times (T - m) \), \( S_{21} = S_{21}^T \) is \( T - m \times m \) and \( S_{22} \) 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., \( S_{11} \) contains the covariance of the measurements actually observed for that case; \( S_{22} \) the covariance between the observed and 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 \( y \) are estimated by Eq. (5).

\[
\hat{\gamma} = (W^T S^{-1} W)^{-1} W^T S^{-1} X
\]

Where \( X \) is the \( T \times 1 \) vector of means at each time point. We then estimate \( X_2 \) and \( \hat{X} \), by Eq. (6) and (7).

\[
\hat{X} = W \hat{Y} + S_{11}^{-1} (X_1 - W_1 \hat{Y})
\]
Kayaşoğlu and Çevik: The estimation of missing values

\[ \hat{Y} = W_1 Y + S_2 S_1^{-1} (X_2 - W_2 Y) \]

W. 1  1  1
1  1  1  
1  3  3

W. 1  2
1  3  9

\[ X = \begin{bmatrix} 21.6 & 22.5 & 27.1 \\ 9.0 & 9.3 & 10.0 \\ 7.5 & 7.3 & 7.9 \\ 10 & 11.1 & 8.9 \end{bmatrix} \]
\[ X_1 = \begin{bmatrix} 17.8 & 23.5 & 22.0 \\ 22.4 & 24.2 & 28.5 \\ 27.2 & 27.9 & 23.0 \\ 25.8 & 28.5 & 25.5 \\ 27.1 & 27.2 & 27.5 \end{bmatrix} \]

For these data X, 12 months and Y, at different stations.

\[ X = \begin{bmatrix} 18.41 & 19.34 & 20.69 \end{bmatrix} \]

\[ S = \begin{bmatrix} 57.92 & 66.01 & 62.72 \\ 66.01 & 60.75 & 71.94 \\ 62.72 & 71.94 & 71.51 \end{bmatrix} \]

D = P - 1 = 3 - 1 = 2

The time design matrix is

\[ W = \begin{bmatrix} 1 & 1 & 1 \\ 1 & 2 & 1 \\ 1 & 3 & 3 \end{bmatrix} \]

and so the estimated regression coefficients for the data are by Eq. (5)

\[ \hat{Y} = \begin{bmatrix} 17.37 \\ 5.28 \end{bmatrix} \]

These missing data were randomized from each period and then be prepared as shown below with periods "*" representing the missing data points.

\begin{align*}
21.6 & & 22.5 & & 27.1 \\
7.5 & & 7.3 & & 7.9 \\
10.0 & & 11.1 & & 9.9 \\
17.5 & & 20.0 & & 22.0 \\
27.0 & & 27.0 & & 29.0 \\
* & & 28.0 & & * \\
27.1 & & 17.0 & & 19.0 \\
\end{align*}

The water temperature (°C) 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. We also use \(\hat{Y}\) as computed previously.

\[ X_1 = \begin{bmatrix} 25 \\ 26.5 \end{bmatrix} \]

\[ X_2 = \begin{bmatrix} 27.2 \\ 27.0 \end{bmatrix} \]

Results and Discussion

The efficiency of this method depends upon the correlation between the dependent variables (period) and the rate of missing observations. While the estimates of the elements of \(y\) are quite similar in the example under consideration when missing data are imputed than when the original complete data set is used. At the end of studies, estimated values for the incomplete observations were similar to the model for the full completed observations. This results showed that the regression method for the missing observations can be used in similar cases by researchers. If the observation had loosed at beginning of experiment one could have made following steps:

Calculation of mean variable which has got missing observation.

Acceptance of this mean value like observed value.

And then the estimation of missing observation by using regression method.

References


