OnLine Journal of Biological Sciences 1 (7): 678-679, 2001
© Asian Network for Scientific Information 2001

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

[1]G. Tamer KAYAALP and [2]Fatma ÇEVÝK
[1]Faculty of Agriculture, [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 value (3 station x 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 $t_1, t_2, ..., t_T$ 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 multi variate normal (Anderson, 1957; Afifi and Elashoff, 1966; Hocking and Smith, 1968; Hartley and Hocking, 1971).

Weighted least squares analyses described by Grizzle *et al.* (1969). They have been developed the analysis of incomplete longitudinal categorical data. Haitovsky (1968) 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, 1975; Donner and Rosner, 1982).

Materials and Methods

30 water temperature values (3 station x 10 water temperature at different months) were taking into account from three station at Seyhan River in Adana – Turkey. It was assumed that each row, X' , of X has a multi variate normal distribution with mean or expected value,

$$E(X) = W\gamma \qquad (1)$$

And (arbitrary) covariance matrix, Σ, is the P x T within-individual (or time) design matrix used to fit a polynomial of degree D = P – 1 to data and γ is the P x 1 vector of polynomial regression coefficients.

$$W = \begin{bmatrix} 1 & t_1 & t_1^2 \\ 1 & t_2 & t_2^2 \\ \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot \\ 1 & t_T & t_T^2 \end{bmatrix} \quad \hat{\gamma} \begin{bmatrix} \gamma_1 \\ \gamma_2 \\ \gamma_P \end{bmatrix} \qquad (2)$$

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

$$X = \begin{bmatrix} X_1 \\ X_2 \end{bmatrix} = \begin{bmatrix} W_1 \\ W_2 \end{bmatrix} \gamma \qquad (3)$$

Where X_1 is (T – m) X_1, X_2 is m x 1, W_1 is (T – m) x P and W_2 is m x 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 x T sample covariance matrix, S , as in equation 4.

$$S = \begin{bmatrix} S_{11} & S_{12} \\ S_{21} & S_{22} \end{bmatrix} \qquad (4)$$

So that S_{11} is (T – m) x (T – m), $S_{12} = S_{21}'$ is (T– m) x m and S_{22} is m x 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_{12} 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 γ are estimated by Eq. (5).

$$\hat{\gamma} = (W' S^{-1} W)^{-1} W' S^{-1} X \qquad (5)$$

Where X is the T x 1 vector of means at each time point. We then estimate X_2 and X_1 by Eq. (6) and (7).

$$\hat{X}_2 = W_2\gamma + S_{21} S_{11}^{-1} (X_1 - W_1 \gamma) \qquad (6)$$

$$\hat{X} = W_1 \gamma + S_{22} S_{12}^{-1} (X_2 - W_2 \gamma) \quad (7)$$

Example
The data is reproduced below (t = 1, 2, 3; P = 3; N = 10).
Where,
t: The number of station
P: The number of regression coefficient
N: The number of observations (water temperature at different months) at station.

$$X = \begin{bmatrix} 21.5 & 22.5 & 27.1 \\ 9.0 & 9.3 & 10.0 \\ 7.0 & 7.3 & 7.9 \\ 10 & 11.1 & 9.9 \\ 17.5 & 20.0 & 22.0 \\ 22.4 & 24.0 & 28.5 \\ 27.0 & 27.0 & 29.0 \\ 25.0 & 28.0 & 26.5 \\ 27.7 & 27.2 & 27.0 \\ 17.0 & 17.0 & 19.0 \end{bmatrix}$$

For these data we found

$\bar{X}$ = (18.41 19.34 20.69)

$$S = \begin{bmatrix} 57.92 & 66.01 & 69.72 \\ 66.01 & 80.75 & 71.98 \\ 69.72 & 71.98 & 71.51 \end{bmatrix}$$

D = P – 1 = 3 –1 = 2

The time design matrix is

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

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

$$\hat{\gamma} = \begin{bmatrix} 17.90 \\ 2.30 \\ 0.21 \end{bmatrix}$$

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

21.5	22.5	27.1
9.0	9.3	10.0
7.0	7.3	7.9
10.0	11.1	9.9
17.5	20.0	22.0
22.4	24.0	28.5
27.0	27.0	29.0
*	28.0	*
27.7	*	*
17.0	17.0	19.0

The water temperature (^{0} 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 γ as computed previously.

$$X_1 = \begin{bmatrix} 28 \\ 26.5 \end{bmatrix} \qquad X_2 = \begin{bmatrix} 27.2 \\ 27.0 \end{bmatrix}$$

$$W_1 = \begin{bmatrix} 1 & 1 & 1 \\ 1 & 3 & 9 \end{bmatrix} \qquad W_2 = \begin{bmatrix} 1 & 2 & 4 \\ 1 & 3 & 9 \end{bmatrix}$$

$$S_{11} = \begin{bmatrix} 57.92 & 69.72 \\ 69.72 & 71.51 \end{bmatrix} \qquad S_{12} = \begin{bmatrix} 66.01 & 69.72 \\ 71.98 & 71.51 \end{bmatrix}$$

$$S_{21} = \begin{bmatrix} 62.75 & 71.98 \\ 71.98 & 71.51 \end{bmatrix}$$

And then the estimated X_2 and X_1 by using Eq. (6) and (7).

$$X_2 = \begin{bmatrix} 25.772 \\ 25.618 \end{bmatrix}$$

$$X_1 = \begin{bmatrix} 28.557 \\ 27.000 \end{bmatrix}$$

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 γ 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

Afifi, A.A. and R. M. Elashoff, 1966. Missing Observations in Multi variate Statistics –I. Review of The Literature. JASA, 61:595-604.

Anderson, T.W., 1957. Maximum Likelihood Estimates For A Multi variate Normal Distribution. When Some Observations Are Missing. JASA, 52: 200-203.

Beale, E.M.L. and R. J. A. Little, 1975. Missing Values in Multi variate Analysis. Journal of The Royal Statistical Society Ser., 37:129-146.

Donner, A. and B. Rosner, 1982. Missing Value Problems in Multiple Linear Regression With Two Independent Variables. Communications in Statistics, 11:127-140

Grizzle, J.E., C. F. Starmer and G. G. Koch, 1969. Analysis of Categorical Data by Linear Models. Biometrics, 25: 489-504.

Hartley, H.O. and R. R. Hocking, 1971. The Analysis of Incomplete Data. Biometrics, 27: 783-808.

Hocking, R. R. and W. B. Smith, 1968. Estimation Of Parameters in The Multi variate Normal Distribution With Missing Observations. JASA, 63: 159-173.

Kayaalp, G.T., 1999. Linear Regression Analysis With Missing Observations Among The Independent Variables in Animal Breeding. Turkish Journal of Veterinary And Animal Sci., 23 : 149-151.

Rubin, D.B., 1976. Inference and Missing Data. Biometrika 63:581-592.

Timm, N.H., 1970. The Estimation of Variance Covariance and Correlation Matrices From Incomplete Data. Psychometrika, 35:417-437