The Estimation of Missing Data With Maximum Likelihood Methods In Animal Research

G. Tamer KAYAALP[1]
E-mail :tamer.kayaalp@gmail.com

[1]Cukurova University Agricultural Faculty Department of Animal Science Biometry and Genetic Section, Turkey

Abstract

The animals probably die in animal experiments. Hence, data were missed at this researches. This situation was produced most problems at experiment.

In this study first step multiple regression analysis was applied for full data. Second step we eliminated randomly 5 % of each independent variables. And then missing data were estimated by using Maximum Likelihood Method.

Keywords: Missing observation, Maximum Likelihood method, coefficient of determination.

Introduction

The study of missing values is one of the most important topics for regression analysis in animal researches. The observations were completed at beginning of experimental. But some experimental materials might be missed during experimental period. For example, in fattening performance some of the animals may die during the experiment because of environmental factors. In this situation, the analysis of experiment might be difficult (İkiz and et. al., 1996).

Anderson (1957) explained that parameters are estimated by using Maximum Likelihood method for data that has multiple variable normal distribution, in case some data are missed at experiment.

Buck (1960) estimated missing observations by improving a method that he gave his name. The researcher's results are similar to Least square method's results.

Glasser (1964) predicted parameters in the linear regression model belong to data set some of which are eliminated randomly in dependent variable. The researcher used Least square method in his study.

Afifi and Elashoff (1966) examined to predict estimators of regression equation which belong to missing observations and statistical properties of its different estimators.

The parameters were estimated by using Least Squares Method and modified Least Square method, in case X and Y variables have got missing observations. Residual mean of squares was used such as compared criteria value in this study (Afifi and Elashoff, 1966).

Haitovsky (1968) illustrated missing observations in artificial data derived via Monte Carlo simulation method for linear regression model by using Least square method and variance-covariance matrix method. Furthermore, the researcher emphasized that Least square method ineffective to estimate regression model's parameters when up to 15 % of all data disappears during experiment period.

Morison (1971) predicted missing observations at independent variable by using Maximum likelihood method and indicated both expected value and variance for Maximum likelihood estimators.

Beale and Little (1975) examined comparatively Maximum likelihood method and Buck's method for estimating the covariance matrix, which has quite stable performance through different patterns in a simulation study and suggested the former method.

Dempster et al. (1977) improved the EM algorithm to cope with missing observations, which are differ proportions, This algorithm is an iterative computational method to get a maximum-likelihood estimate when the data can be conveniently viewed as incomplete.

Donner (1982) predicted parameters of linear regression model, which have two variables (X_1 and X_2), for all data and missing observations.

Shih and Weisberg (1986) improved an algorithm that belongs to Maximum likelihood method and predicted parameters of linear regression model by using this algorithm. Here, Cook's statistics was used as compared test statistics.

Simonoff (1986) explained factors that influence the estimation, in case missing observation values are increased in data set.

Kayaalp (1999) has illustrated a method for multiple regression models with missing data. 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 distribution (Anderson, 1957; Kayaalp and Çevik, 2000).

Atkinson and Cheng (2000) explained that The combinations of the forward search algorithm with the EM algorithm or multiple imputation is successful in detecting outliers from linear regressions models with an appreciable proportion of missing values and also provides a very robust estimation procedure.

The aim of this study is that regression equations are estimated by using maximum likelihood method for full and incomplete observations respectively. And then we are going to illustrate of estimating missing observations at experiment.

Material and Method

Material

In this study, used data, which belong to fattening performance, are recorded in Cukurova University, Agricultural Faculty, Animal Science Department, Feeds and Animal Nutrition Unit. There are sixty data for each variables used in analysis. Each variable's symbols has illustrated as following,

Y: Carcass weight
X_1: Body weight gain
X_2: Thigh weight

X_3: Forearm weight

Method

In this study, multiple regression model was used and its model is given in the Equation (1).

$$Y_i = a + b_1 X_1 + b_2 X_2 + b_3 X_3 + e_i \quad (1)$$

Where,

Y_i : Dependent variable,

X_i : Independent variable,

a : This parameter must then be the value of Y when X is equal to zero (or intercept).

b_i : Regression coefficient (or slope)

e_i : Residuals ~ N (0, $\sigma^2 e$)

This model's parameters were estimated by using MINITAB statistic package program 13.1 Version.

Here, initially, the regression's parameters were estimated for all of observations. After that, five percent ratio of data of each X variables was randomly lost. The last process, multiple regression model's parameters were estimated by using Maximum Likelihood method. To obtain missing observations, three observations of each dependent variable were eliminated randomly.

Maximum Likelihood Method

1) Variance – covariance matrix is produced as following:

$$\sigma_{ij} = \begin{bmatrix} \sigma_{X_1X_1} & \sigma_{X_1X_2} & \sigma_{X_1X_3} \\ \sigma_{X_2X_1} & \sigma_{X_2X_2} & \sigma_{X_2X_3} \\ \sigma_{X_3X_1} & \sigma_{X_3X_2} & \sigma_{X_3X_3} \end{bmatrix}$$

The values of variance – covariance matrix are calculated from Equations (2), (3), (4), (5), (6), (7), (8) respectively.

$$\sigma_{X_1X_1} = \sum X_1^2 - (\sum X_1)^2 / n \quad (2)$$

$$\sigma_{X_2X_2} = \sum X_2^2 - (\sum X_2)^2 / n \quad (3)$$

$$\sigma_{X_3X_3} = \sum X_3^2 - (\sum X_3)^2 / n \quad (4)$$

$$\sigma_{X_1X_2} = \sum X_1 X_2 W_1 W_2 - (\sum X_1 W_1)(\sum X_2 W_2) / n \quad (5)$$

$$\sigma_{X_1X_3} = \sum X_1X_3W_1W_3 - (\sum X_1W_1)(\sum X_3W_3)/n \quad (6)$$

$$\sigma_{X_2X_3} = \sum X_2X_3W_2W_3 - (\sum X_2W_2)(\sum X_3W_3)/n \quad (7)$$

If data were lost $W_i = 0$, else $W_i = 1$.

2) The covariance vector of Y variable is produced where covariance between X and Y variable is computed.

$$\sigma_{ij} = \begin{bmatrix} \sigma_{X_1Y} \\ \sigma_{X_2Y} \\ \sigma_{X_3Y} \end{bmatrix}$$

The covariance of this vector are calculated from Equations (8), (9), (10) respectively.

$$\sigma_{X_1Y} = \sum X_1Y W_1 - (\sum X_1W_1)(\sum Y)/n \quad (8)$$

$$\sigma_{X_2Y} = \sum X_2Y W_2 - (\sum X_2W_2)(\sum Y)/n \quad (9)$$

$$\sigma_{X_3Y} = \sum X_3Y W_3 - (\sum X_3W_3)(\sum Y)/n \quad (10)$$

3) The vector $\hat{b}' = (\hat{b}_1, \hat{b}_2, \hat{b}_3)$ is estimated from equation (11).

$$b' = (\sigma_{ij})^{-1}\sigma_{iY} \quad (11)$$

4) a value is estimated from equation (12).

$$\hat{a} = \overline{Y} - b_1\overline{X}_1 - b_2\overline{X}_2 - b_3\overline{X}_3 \quad (12)$$

5) The missing observations are estimated by using predicted regression equation.

6) Multiple regression analysis is completed by using Maximum Likelihood method, after missing observations' values are replaced in the analysis progress.

The regression analysis' table was given in Table 1 in order to test of significance for regression equation.

Table 1. Regression Analysis Table.

S.V.	**D.F.**	**S.S**	**M.S.**	**F**
Regression	P	$b_1S_{x_1y} + b_2S_{x_2y} + b_3S_{x_3y} = A$	$RMS = A/p$	RMS/MSE
Error	N-p-1	$SSE = S_{yy} - A$	$MSE = SSE/(N-p-1)$	
Total	N-1	S_{yy}		

Here,

p : The number of independent variable's value,

N : The number of observation value,
RMS : Means of square of regression
SSE : Sum of square of error,
MSE : Means of sum of square of error,

Coefficient of Determination (R^2), which belongs to predicted multiple regression model in this study, can be computed by using Equation (13),

$$R^2 = \frac{b_1 S_{x_1 y} + b_2 S_{x_2 y} + b_3 S_{x_3 y}}{S_{yy}} \quad (13)$$

Results

When there are all of data, by using Equation (1), multiple regression model's parameters were predicted and given as following equation,

Carcass weight (Y) = -3.16 + 5.67 Body weight gain (X_1) + 0.00626 Thigh weight (X_2) + 0.00207 Forearm weight (X_3)

In the model, although the first two parameters of the regression coefficients, b_1 and b_2, were statistically significant ($p<0.01$), last parameter, b_3, was not significant ($p>0.01$). The R^2 of predicted regression model was computed as 0.849.

The test of significance for predicted regression equation was given in Table 2.

Table 2. Regression Analysis Results for all data.

S.V	D.F.	S.S	M.S.	F
Regression	3	210.244	70.081	104.58**
Error	56	37.527	0.670	
Total	59	247.771		

**: $p<0.01$

Table 2 shows that predicted regression model is statistically significant ($p<0.01$). In addition, a parameter of the model is statistically significant, too.

When there are missing observations, by using Maximum Likelihood method, predicted results of multiple regression model's parameters were given as following.

Variances and covariances belong to X and Y variables were computed from Equation (2), (3), (4), (5), (6), (7), (8), (9), (10) respectively.

$$\sigma_{X_1X_1} = 0.002; \quad \sigma_{X_2X_2} = 66011.601; \quad \sigma_{X_3X_3} = 18066.060;$$
$$\sigma_{X_1X_2} = 4.289; \quad \sigma_{X_1X_3} = 1.781; \quad \sigma_{X_2X_3} = 28160.944;$$
$$\sigma_{X_1Y} = 0.0393; \quad \sigma_{X_2Y} = 485.624; \quad \sigma_{X_3Y} = 225.440$$

The vector, $\hat{b}' = (\hat{b}_1, \hat{b}_2, \hat{b}_3)$, was estimated by using Equation 11.

$$\hat{b}' = (\sigma_{ij})^{-1}\sigma_{X_iY} = \begin{bmatrix} 0.002 & 4.289 & 1.781 \\ 4.289 & 66011.601 & 28160.944 \\ 1.781 & 28160.944 & 18066.060 \end{bmatrix}^{-1} \begin{bmatrix} 0.0393 \\ 485.624 \\ 225.440 \end{bmatrix}$$

$$\hat{b}' = (\sigma_{ij})^{-1}\sigma_{X_iY} = (4.60167, \quad 0.00575, \quad 0.00306$$

a value was estimated from equation (12) as following,

$$\hat{a} = \overline{Y} - b_1\overline{X}_1 - b_2\overline{X}_2 - b_3\overline{X}_3 = -2.801$$

As a result, by using Maximum Likelihood method, when the parameters were replaced in the regression equation, the equation was determined as following equation,

Carcass weight (Y) = -2.801 + 4.60167 Body weight gain (X_1) + 0.00575 Thigh weight (X_2) + 0.00306 Forearm weight (X_3)

Determination coefficient (R^2) was computed as 0.871. In additional to the results, in this model, all of the parameters were statistically significant ($p<0.01$).

The test of significance for predicted regression equation was given in Table 3.

Table 3. By using Maximum Likelihood Method, Regression Analysis Results for data some of which are missed

S.V	**D.F.**	**S.S**	**M.S.**	**F**
Regression	3	184.225	61.408	105.88**
Error	47	27.227	0.580	
Total	50	211.452		

** : $p<0.01$

In Table 4, missing observations given in parentheses were estimated by using predicted regression model.

Table 4. By Using Maximum Likelihood Method Missing Values predicted

Body Weight Gain (X_1)		Thigh weight (X_2)		Forearm weight (X_3)	
Observed Values	Expected Values	Observed Values	Expected Values	Observed Values	Expected Values
0.32 (16)	0.36	2778 (15)	2820.18	1213	1053.92
0.35 (18)	0.37	2412 (31)	2630.09	(7)	985.38
0.34 (55)	0.52	3090 (49)	3279.34	1303 (38)	2151.31
				1290 (56)	

Discussion

Table 2 shows that predicted regression model was found significant for all data ($p< 0.01$). According to the results, the carcass weight may be estimated in predicted regression model by replacing body weight gain (X_1), thigh weight (X_2) and forearm weight (X_3). In case, there are missing observations, the derivation between estimated parameters and original

parameters no much in predicted regression model for all data. Even some observations are missed during experiment period, missing observation can be estimated by using predicted regression model. Table 3 showed that the derivation between estimated values and original values no much in our study expect for body weight gain belong to 55. observation value, because the correlation coefficient is much between body weight gain and carcass weight as 0.78.

Maximum Likelihood method's activity is dependent to missing observation's proportion and to correlation value between independent variable (X) and dependent variable (Y). In general, if missing observations are up to 15 % of all data and correlation value between X and Y variables is up to 0.70, this method's activity is decreased and the derivation between observed and expected values will be increased.

References

İkiz, F., Püskülcü, H. ve Eren, Ş., 1996. İstatistiğe Giriş. İzmir, 435p.

Anderson, T.W., 1957. Maximum Likelihood Estimate for a Multivariate Normal Distribution When Some Observations are Missing. Jo Amer. Statist. Assoc. (52): 200-204.

Buck, S.F., 1960. A Method of Estimation of Missing Values in Multivariate Data Suitable for Use with An electronic Computer. J.Roy.Statist. Soc. (B22): 302-307.

Glasser, M., 1964. Linear Regression Analysis with Missing Observations Among The Independent Variables. J. Amer. Statist. Assoc. (59): 834-844.

Afifi, A.A. and Elashoff, R.M., 1966. Missing Observations in Multivariate Statistics I. Review of The Literature. J. Amer. Statist. Assoc. (61):595-604.

Haitovsky, Y., 1968. Missing Data in Regression Analysis. J. Amer. Statist. Assoc. (59): 834-844.

Morrison, D.F., 1971. Expectations and Variances of Maximum Likelihood Estimates of The Multivariate Normal Distribution Parameters with Missing Data. J. Amer. Statist. Assoc. (66): 602-604.

Beale, E.M.L. and Little R.J.A., 1975. Missing Values in Multivariate Analysis. J. Roy. Statist. Soc. Ser. (B37):129-146.

Dempster, A., Laird, N.M., Rubin, D.B., 1977. Maximum Likelihood Estimation From Incomplete Data Via The EM Algorithm. J. Roy. Statist. Soc. Ser. (B39):1-38.

Donner, A., 1982. The Relative Effectiveness of Procedures Commonly Used in Multiple Linear Regression Analysis for Dealing with Missing Values. The Amer. Statist. Assoc. (38):378-381.

Shih, W.J. and Weisberg, S., 1986. Assessing Influence in Multiple Linear Regression with Incomplete Data. Technometrics. (28): 231-239.

Simonoff, J.S., 1986. Regression Diagnostics to Detect Nonrandom Missingness in Linear Regression. Technometrics. (30): 205-214.

Kayaalp, G.T., 1999. Linear Regression Analysis with Missing Observations among The Independent Variables in Animal Breeding. TUBITAK Turkish J. Vet. and Anim. Sci. **(2): 149-151.**

Kayaalp, G.T. and Çevik F., 2000. The Estimation of Missing Values in Longitudinal Data Sets by Using Regression Methods in Biological Researches. Online Journal of Biological Sciences. (7): 678-679.

Atkinson, A:C and Cheng, T., 2000. On Robust Linear Regression with Incomplete Data. Computational Statistics & Data Analysis. (33): 361-380.