Tr. J. of Veterinary and Animal Sciences
23 (1999) 149-151

Linear Regression Analysis With Missing Observations Among The Independent Variables in Animal Breeding

G. Tamer KAYAALP
Department of Animal Science, Faculty of Agriculture, University of Gaziosmanpaşa, Tokat-TURKEY

Received: 10.10.1997

Abstract: In animal breeding, when there is a relationship between the dependent (Y) and independent (X) variables, regression analysis is applied. But when one of the variables has one or more missing observations regression analysis cannot be applied. This paper illustrates and discusses a regression analysis in which the independent variable (X) has a missing observation.

Key Words: Regression Analysis, Least Square Estimation, Missing Observations.

Hayvancılıkta Bağımsız Değişkenler Arasında Kayıp Gözlemli Doğrusal Regresyon Analizi

Özet: Hayvancılıkta bağımsız (X) değişken ile bağımlı (Y) değişkenler arası ilişki regresyon analizi ile ifade edilir. Fakat bu değişkenlerden birisi bir veya daha fazla kayıp gözleme sahip ise regresyon analizi yapılamaz. Bu makalede X değişkeninde (bağımsız değişken) kayıp gözlem olduğunda regresyon analizinin uygulanışı tanıtılmış ve tartışılmıştır.

Anahtar Sözcükler: Regresyon Analizi, En Küçük Kareler Tahmini, Kayıp Gözlemler.

Introduction

A pair of random variables such as (height and weight) follows some sort of bivariate probability distribution. When we are concerned with the dependence of a random variable Y on a quantity X which is not a random variable, an equation that relates Y to X is usually called a regression equation (1).

Regression analysis has been loosely described as a study of the relationship between one variable, the response variable, and one or more other variables, the predictor variables. The parameters of regression (b_0, b_1) were estimated by the least square technique (2).

Dear (2), has illustrated a method for multiple regression models with missing data.

Method

It was assumed that we had a sample of size N from a (p + 1) multivariate distribution as:

$$f(Y \mid x_1, x_2, \ldots, x_p) \sim N(\beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p, \sigma^2 e).$$

The maximum likelihood solution for the b's in the classical regression situtation, where the X's are considered to be fixed with no missing observations may be stated as follows:

$$\hat{\beta}_j = \sum_{k=1}^{p} (\hat{\sigma}_{jk})^{-1} \hat{\sigma}_{k0} \qquad (j \neq k) \tag{1}$$

$$\hat{\beta}_0 = \bar{Y} - \beta_j \bar{X}_j$$

$$\hat{\sigma}_{k0} = \Sigma (X_{ij} - \bar{X}_j)(Y_i - \bar{Y}) / (N - 1)$$

$$\hat{\sigma}_{jk} = \Sigma (X_{ij} - \bar{X}_j)(X_{ik} - \bar{X}_k) / (N - 1) \tag{2}$$

$$(\hat{\sigma}_{jk})^{-1} = 1 / \hat{\sigma}_{jk} \qquad (j, k \neq 0)$$

x_{ij} : value of X_j for the ith individual

$$\bar{X}_j = \Sigma X_{ij} / N, \quad \bar{Y} = \Sigma Y_i / N$$

Now let us introduce missing observations among the fixed independent variables for certain individuals.

We define a random indicator function W_{ij}, such as

Table 1. Regression Analysis With No Missing Observations.

Source of Var.	Degree of Freedom (D.F.)	Sum of Squares (S.S.).
Regression	p	$\Sigma S_{xiy} \beta_i$ [2]
Residual	n – p – 1	[1] – [2]
Total	n – 1	S_{yy} [1]

p : The number of independent variables.
n : The number of observations.

Table 2. Regression Analysis With Missing Observations.

Source of Var.	D.F.	S.S.
Regression	p	$W_{ij} \Sigma S_{xjy} \beta_j$ [2]
Residual	n – 1 – k – p	[1] – [2]
Total	n – 1 – k	S_{yy} [1]

k : The number of missing observations.

Example

The food consumption of six chickens (Y) (kg/350 day) and body weight (X) (kg) are given in Table 3.

Table 3. The Food Consumption Six Chickens (Y) (kg/350 day) Body Weight (X) (kg) (4).

X	2.08	2.31	2.17	1.99	2.67	2.13
Y	39.45	42.17	40.68	41.40	45.07	41.72

Regression parameters (β_0 and β_1) were estimated from the data in table 3.

Case 1: With no missing observations.

Regression parameters (β_0 and β_1) were estimated as described in equation 1 and Table 1.

$\hat{\beta}_0 = 26.661$

$\hat{\beta}_1 = 6.781$

Table 4. Regression Analysis With No Missing Observations.

S.V.	D.F.	S.S.	M.S.	F
Regression	1	13.498	13.498	12.674
Residual	4	4.259	1.065	
Total	5	17.757		

* : $p < 0.05$

Case 2: With missing observations.

We assumed that X's fourth observation value in Table 3 (1.99) was missing. Then we estimated the regression parameters.

X	2.08	2.31	2.17	*	2.67	2.13
Y	39.45	42.17	40.68	41.40	45.07	41.72
W_{ij}	1	1	1	0	1	1

* : Missing observation.

$W_{ij} = 1$ if X_{ij} is observed.

$W_{ij} = 0$ if X_{ij} is not observed.

Randomness of the missing observations means that the joint distribution of any set of W's is given as the product of the probabilities of the individual indicator functions.

$N_j = \Sigma W_{ij}$

(The number of individuals for which X_j is observed).

$N_{jk} = \Sigma W_{ij} W_{ik}$

(The number of individuals for which both X_j and X_k and observed).

$\bar{X}_{j\,(j)} = \Sigma W_{ij} X_{ij} / N_j$

(The mean of X_j's based on observed values of X_j's for individuals for which both X_j and X_k are observed).

$\bar{Y}_{(j)} = \Sigma W_{ij} Y_i / N_j$

(The mean of Y's for individuals for which X_j is observed).

By analogy, it would appear reasonable to apply estimates similar to (1) by using all available data.

$$\hat{\sigma}_{jk} = \Sigma W_{ij} W_{ik} \left(X_{ij} - \bar{X}_{j\,(jk)}\right)\left(X_{ik} - \bar{X}_{k\,(jk)}\right)/(N_{jk} - 1) \quad (j, k \neq 0)$$

$$\hat{\sigma}_{j0} = \Sigma W_{ij} \left(X_{ij} - \bar{X}_{j\,(j)}\right)\left(Y_i - \bar{Y}_{(j)}\right) / \left(N_j - 1\right)$$

Our estimates may now be written:

$$\hat{\beta}_j = \Sigma (\hat{\sigma}_{jk})^{-1} \hat{\sigma}_{k0} \qquad (j \neq 0)$$

$$\hat{\beta}_0 = \bar{Y} - \Sigma \beta_j \bar{X}_{j\,(j)} \qquad (3)$$

$N_1 = \Sigma W_{i1} = 5$

$\bar{X}_{1(1)} = \Sigma W_{i1} X_{i1} / N_1 = \left((1)(2.08) + \ldots + (1)(2.13)\right)/5 = 2.272$

$$\bar{Y}_{(1)}=\Sigma W_{i1}Y_i/N_1=(1)(39.45)+...+(1)(41.72)/5=41.818$$

$$\hat{\sigma}_{11} = \Sigma W_{i1} W_{i1} \left(X_{i1} - \bar{X}_{1(1)}\right)^2 / \left(N_1 - 1\right)$$

$$= \left((1) (1) (2.08 - 2.272)\right)^2 + ... +$$

$$(1) (1) \left(2.13 - 2.72\right)^2 / (5 - 1) = 0.05682$$

$$\hat{\sigma}_{10} = \Sigma W_{i1} \left(X_{i1} - \bar{X}_{1(1)}\right) \left(Y_i - \bar{Y}_{(1)}\right) / \left(N_1 - 1\right)$$

$$= [(1) (2.08 - 2.272) (39.45 - 41.818) + ... + (1)$$

$$(2.13 - 2.272) (41.72 - 41.818)] / 4 = 0.47308$$

$$\hat{\beta}_1 = \Sigma \left(\hat{\sigma}_{11}\right)^{-1} \hat{\sigma}_{10} = (1 / 0.05682) (0.47308) = 8.326$$

$$\hat{\beta}_0 = \bar{Y} - \beta_1 \bar{X}_{1\ (1)} = 41.748 - (8.326)(2.272) = 22.831$$

Table 5. Regression Analysis With Missing Observations.

S.V.	D.F.	S.S.	M.S.	F
Regression	1	15.755	15.755	23.621*
Residual	3	2.000	0.667	
Total	4	17.757		

* : $p < 0.05$.

Discussion and Results

These estimates are consistent estimates of the usual unbiased estimates as can be seen from the following considerations:

1) The X_{ij}, X_{ik} pairs of values are from a finite population from which members are chosen independently.

2) The σ_{jk} (j, k ≠ 0) are, therefore, random variables and are the usual unbiased estimates of the covariance for a finite population.

3) The σ_{jk} are obviously consistent estimates of σ_{jk} since their variances approach 0.

4) The same reasoning can be applied to σ_{k0}.

5) Similarly the $X_{j(j)}$ are consistent estimates of X_j.

6) The efficiency of this method depends upon the correlation between the independent variables and the rate of missing observations.

The results of this study are as follows:

1) Regression was significant ($P < 0.05$) as shown in tables 4 and 5.

2) X's 4th observation value was predicted to be 2.230.

References

1. Draper, N.R. and Smith, H., Applied Regression Analysis. NewYork John Wiley and Sons Inc. 1996.
2. Glasser, M., Linear Regression Analysis With Missing Observations Among The Independent Variable. JASA Vol: 59 834-844. 1964.
3. Dear, R.E., Principal Component Missing Data Method For Multiple Regression Models. SP-86 System Development Corporation, Santa Monica, California, 1959.
(4) İkiz, F., Püskülcü, H. ve Eren, Ş. İstatistiğe Giriş. 435 Sayfa. 1996.