ISSN:0971-1260
Vol-22- Issue-14-December-2019

# Estimation Of Parameters In The Presence Of Missing Data In Errors In Variables Linear Models: A New Estimator

**D. Kavitha* & V.B.Naidu****

*Research Scholar, Department of Econometrics, S.V.University, Tirupati, A.P., India
**Professor (Retd.), Department of Econometrics, S.V.University, Tirupati, A.P., India

**ABSTRACT**

Estimation of parameters of the general linear model will pose problems in the presence of measurement errors with missing data. In some cases there will be gaps in the data or some observations may be completely missing. For example, in an agriculture experiment the observations from some plots may not be available. These situations necessitate the estimation of the missing observations. In these contexts some standard techniques for predicating the observations would be of much use. Considerable attention was paid by several researchers for this particular pattern of missing observations and a lot of literature appeared. In this paper an attempt has been made to provided an estimator of parameter vector in the presence of measurement errors and missing data. Extensive Monte Carlo study revealed that the new estimator is more efficient than OLE.

## 1.1 INTRODUCTION

Estimation of parameters of the general linear model will pose problems in the presence of measurement errors with missing data. There are several simple methods by which one can estimate missing values and then estimate the parameters. However, most of the work on missing observations assumes that all the data may be considered as a random sample from a known multivariate distribution, usually taken to be multi- normal. As most of the econometric analysis is based on the general liner modal with stochastic regressors when the variables are affected by observational errors with missing data, there is a need to have estimators, which are useful for practical applications.

There are variety of reasons why the data may be missing. In some cases the desired series may just be unobservable. In some others, like cross-section data, where data is collected by means of survey, there may be some respondents who may not answer some questions or not

 

ISSN:0971-1260
Vol-22- Issue-14-December-2019

respond at all. In such cases there will be gaps in the data or some observations may be completely missing. For example, in an agriculture experiment the observations from some plots may not be available. These situations necessitate the estimation of the missing observations. In these contexts some standard techniques for predicating the observations would be of much use.

Considerable attention was paid by several researchers for this particular pattern of missing observations and a lot of literature appeared. Hence, the researcher has provided an estimator of parameter vector in the presence of measurement errors and missing data in this thesis.

**1.2 MISSING OF OBSERVATIONS**

There are several simple methods by which one can estimate missing values and then estimate the parameters. However, most of the work on missing observations assumes that all the data may be considered as a random sample from a known multivariate distribution, usually taken to be multi- normal. Commencing with the work of Anderson (1957), a large literature has appeared. Hartley and Hocking (1971) and Rubin (1974) have outlined the pattern of missing values for which explicit algebraic formulae for maximum likelihood solutions are possible. Dumpster, Laird and Rubin (1977) have discussed numerical approaches for the other, less algebraically tractable situations. Gorier and Montfort(1981) have derived M.L. estimators for this same pattern without making any distributional assumptions about the explanatory variables present for all observations and compared with the maximum likelihood approach of Dagenasi (1973).

The problem of missing values in discriminate analysis is discussed by Chas and Dunn (1972, 1974), using the method proposed by Dear (1959). The missing data problem in factor analysis is considered by Woodbury and Saber (1960). Press and Scott (1975, 1976) and Swami and Mehta (1975) used the Bayesian Regression procedure to estimate the missing values and the parameters.

Beale and Little (1975) have applied maximum likelihood method to estimate missing value. They have used the missing information principle suggested by Orchard and Woodbury (1972) in their analysis. Hartley and Hocking (1971) and Rubin (1974) have outlined the

ISSN:0971-1260
Vol-22- Issue-14-December-2019

pattern of missing values for which explicit algebraic formulae for M.L. solutions. Orchard and Woodbury (1972) and Dumpster, Laird and Rubin (1977) have discussed numerical approaches for the other, less algebraically tractable situations.

Denis Connie (1983) proposed other estimators for this pattern of missing data and has shown that the estimators are unbiased in small sample and derived explicitly small sample variance formulae. A.C. Harvey and C.R. Mackenzie (1984**)** have examined a number of methods for carrying out the maximum likelihood estimation of a dynamic econometric model with missing observations.

Maher EI-Masri, Susan M Fox-Wasylyshyn (2005) have provided the novice with an introductory conceptual issue of missing data. Paul D Allision (2012**)** has presented an intuitiveestimate for estimating linear regression coefficients when an unspecified number and pattern of the independent variables for some individuals are not observed. Mohd Mustafa Al and Bakri Abdullah (2014)havediscussed three interpolation methods that are linear, quadratic and cubic. Blackwell and James (2017) have developed a unified approach to measurement errors and missing data.

In (1972), Orchard and Woodbury have presented a method of estimation of missing observations in a general linear model set up by providing a missing information principle. The equations presented by them can be iteratively used to estimate the parameters. The method is briefly outlined.

Consider the case of missing observations in a linear model,

$$Y = X\beta + \in \qquad (1.2.1)$$

Suppose there are "n" potential observations but "m" are missing. Also the observed values are unaltered while the missing values are replaced by their expected values.

Let $X_m$ and $X_0$ are $n \times k$ matrices such that $X = X_m + X_0$ and $Y_m$ $Y_0$ are $(n \times 1)$ vectors such that $Y = Y_m + Y_0$

ISSN:0971-1260
Vol-22- Issue-14-December-2019

Where "*m*" represents missing value and "*o*" observed value of the variables respectively. In such a situation the missing information principle is used which employees the following equations.

$$\hat{\beta} = \left(X^1 X\right)^{-1} X^1 \hat{Y} \qquad (1.2.2)$$

$$\hat{Y} = Y_0 + X_m \hat{\beta} \qquad (1.2.3)$$

And

$$\hat{Y}_m = \left[I - X_m \left(X^1 X\right)^{-1} X_m^1\right]^{-1} X_m \left(X^1 X\right)^{-1} X_0^1 Y_0 \qquad (1.2.4)$$

Equation (4) is used to obtain an initial set of estimates for the missing values, and then the iterative procedure is used until the parameter estimates are stabilized.

The covariance matrix for $\hat{\beta}$ may be written as

$$\text{Cov}\left(\hat{\beta}\right) = \left\{\left(X^1 X\right)^{-1} + \left(X^1 X\right)^{-1} X_m^1 \left[I - X_m \left(X^1 X\right)^{-1}\right]^{-1} X_m \left(X^1 X\right)^{-1}\right\} \sigma^2 \qquad (1.2.5)$$

The main reason for using the above procedure is that the design matrix X for the complete data, where we have a balance design, may be chosen such that $\left(X^1 X\right)$ is diagonal, and hence much easier to invert than the general matrix $\left(X_0^1 X_0\right)$ which would result from using the available data only.

## 1.3 NEW ESTIMATOR

Let the model with stochastic regressors

$$Y = X\beta + \in$$

Corresponding to the true model, the observed model be

$$Y^* = X^* \beta^* + \varepsilon^*$$

Where, $Y^* = Y + U$ and $X^* = X + Z$

The OLS of $\hat{\beta}^*$ is given by

$$\hat{\beta}^* = (X^{1*} X^*)^{-1} X^{1*} Y^*$$

But $\hat{\beta}^*$ is not a consistent of β

ISSN:0971-1260
Vol-22- Issue-14-December-2019

Therefore, the following estimator can be used, which is consistent and it is given by

$$\tilde{\beta}^* = [X^{*1}X^* - W]^{-1} X^{*1}Y^*$$

Where, $W = X^1Z + Z^1X + Z^1Z$

If W is known, then it can be estimated by Extended Ridge method proposed by K.N Lankipalli and et al (1985).

When there are observational errors and missing values, then the new estimator proposed to estimate the parameters of the model is

$$\tilde{\beta}^* = [X^{*1}X^* - W]^{-1} X^{*1}\hat{Y}^*$$

Where, $\hat{Y}^* = Y_0^* + X_m^* \tilde{\beta}^*$

It is shown that $\hat{\beta}^*$ is consistent and asymptotically unbiased for β. The asymptotic variance of $\tilde{\beta}$ is also derived.

## 1.4 MONTE CARLO STUDY

To study the efficiency of the new estimator where there are measurement errors with missing of observations in linear model the following three variable linear model is considered as

$$Y = \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \varepsilon \qquad (1.4.1)$$

Where $Y^* = Y + U$

$$X_j^* = X_j + Z_j \quad \text{j= 1,2,3} \qquad (1.4.2)$$

Where Y and $X_j$ are the true values, Y* and $X_j^*$ are the measured values and U and $Z_j$ are the errors. The efficiency of the new estimator is examined in the following five cases.

Case 1 $X_j \sim N(0,1)$
$Z_j \sim N(0, 0.1)$ j=1, 2, 3
$U \sim N(0, 0.1)$

Case 2 $X_j \sim N(0, 1)$
$Z_j \sim L_n(0, 0.1)$ j=1, 2, 3

ISSN:0971-1260
Vol-22- Issue-14-December-2019

$U \sim L_n$ (0, 0.1)

Case 3 $X_j \sim N$ (0, 10)

$Z_j \sim E$ (0, 0.1) j=1, 2, 3

$U \sim E$ (0, 0.1)

Further in each of the above three cases, the following three different parameter structures are considered:

| | | | | |
|---|---|---|---|---|
| (i) | $\beta_1 = 0.01$ | $\beta_2 = 0.2$ | $\beta_3 = 0.8$ | (1.4.3) |
| (ii) | $\beta_1 = 0.1$ | $\beta_2 = 2.0$ | $\beta_3 = 5.0$ | (1.4.4) |
| (iii) | $\beta_1 = -0.5$ | $\beta_2 = 5.0$ | $\beta_3 = 10.0$ | (1.4.5) |

With the parameter structure (1.4.3) to (1.4.5) and the observations generated on $X_j$ (j =1, 2, 3), the observations on the dependent variable Y are obtained by using the relation (1.4.2). Errors $Z_j$ and u are introduced in to the variables $X_j$ and Y according to the relation (1.4.2)

The process of takingrandom samples on the variables $X_j, Z_j$ (j =1, 2, 3), that are uniform in the interval (0,1) are generated. 50 sets of samples sizes 10,20, 30, 50 and 100 are considered in computing the estimator in each of the three different cases for three structures.

The new estimator is

$$\tilde{\beta}^* = [X^{*1}X^* - W]^{-1} X^{*1}\hat{Y}^*$$

In practice, $\tilde{\beta}^*$ is computed by replacing W by $\hat{W}$, an estimator of W, obtained by Extended Ridge Method. Therefore an “Extended Ridge Method” is developed in the present study to estimate W. When W is estimated by “Extended Ridge Method”, and used to estimate the parameter vector β, it is also difficult to establish the properties of $\tilde{\beta}$. Therefore an extensive Monte Carlo Study is carried out. The experimental study reveals that the new estimator $\tilde{\beta}$ is ‘better’ than OLS when the

 

ISSN:0971-1260
Vol-22- Issue-14-December-2019

true variables and errors follow different distribution viz., Normal, Lognormal and Exponential.

We assume that $W = K_j I$ and $K_j$ assumes positive or negative values according as

$$|X^1 X| \lessgtr |X^{1*} X^*|$$

At each stage of iteration and the iterations are carried out until $\tilde{\beta}_j^*$ is stabilized. In all the three cases measure of efficiency of the new estimator to that of OLE viz.,

$$e = \left|\frac{MSE(\hat{\beta}^*)}{MSE(\tilde{\beta}^*)}\right|$$ is provided.

## 1.5 SUMMARY OF THE RESULTS

From the Monte Carlo study, the following observations are made:

1. The bias and mean square error of the new estimator are smaller than OLE for all the three cases.
2. As the sample size increases, the mean square error of the new estimator constantly decreasing at a faster rate implying that new estimator is more efficient than OLE.
3. For all structures with different combinations of errors the value of e>1, implying that the new estimator is more efficient than OLE.

**TABLE – 1: MEASURE OF EFFICIENCY**

| **CASE** | **SAMPLE SIZE** | **STRUCTURE** | | |
|---|---|---|---|---|
| | | **I** | **II** | **III** |
| | **10** | $10^6(0.139296)$ | $10^7(0.673730)$ | $10^7(0.915690)$ |
| | **20** | $10^7(0.840103)$ | $10^7(0.524964)$ | $10^7(0.407189)$ |

ISSN:0971-1260
Vol-22- Issue-14-December-2019

| | | | | |
|---|---|---|---|---|
| **I** | **30** | $10^{12}$(0.129307) | $10^{8}$(0.543324) | $10^{5}$(0.757546) |
| | **50** | $10^{3}$(0.603703) | $10^{6}$(0.828820) | $10^{5}$(0.277225) |
| | **100** | $10^{7}$(0.773183) | $10^{3}$(0.203448) | $10^{2}$(0.483982) |
| **II** | **10** | $10^{5}$(0.310905) | $10^{5}$(0.634685) | $10^{8}$(0.335418) |
| | **20** | $10^{6}$(0.552619) | $10^{3}$(0.144995) | $10^{2}$(0.079399) |
| | **30** | $10^{6}$(0.441227) | $10^{3}$(0.673392) | $10^{2}$(0.331830) |
| | **50** | $10^{5}$(0.288201) | $10^{4}$(0.166364) | $10^{2}$(0.672898) |
| | **100** | $10^{5}$(0.954594) | $10^{4}$(0.750505) | $10^{3}$(0.651000) |
| **III** | **10** | $10^{4}$(0.324293) | $10^{5}$(0.949970) | $10^{2}$(0.602756) |
| | **20** | $10^{4}$(0.894434) | $10^{2}$(0.611162) | $10^{14}$(0.729343) |
| | **30** | $10^{7}$(0.204872) | $10^{4}$(0.286816) | $10^{2}$(0.873830) |
| | **50** | $10^{4}$(0.337137) | $10^{2}$(0.375331) | $10^{2}$(0.565020) |
| | **100** | $10^{3}$(0.013010) | $10^{3}$(0.313686) | $10^{2}$(0.050000) |

## REFERENCES

1) Alheety, M.I, &Kibria, B.M.G. (2014), “A generalized stochastic restricted ridge regression estimator”, communication in statistics-Theory and Methods, 415-427.

2) Alheety, M.I, &Kibria, B.M.G. (2011), “Choosing ridge parameters in the linear regression model with A.R (1) Error”: A comparative simulation study, Bulletin of Statistics & Economics, p.10-26.

 

ISSN:0971-1260
Vol-22- Issue-14-December-2019

3) A.C. Harvey& Mackenzie, (1984)"Missing observations in dynamic econometric models: A partial synthesis", Springer link, Time series analysis of irregularly observed data, PP 108-133.

4) Alkhamisi, M, Khalaf, G, &Shukur, (2014), "Some modifications for choosing ridge parameter", communication in statistics, Theory and Methods, 43(20).

5) Beale, E.M.L and Little, R.J.A. (1975), "Missing values in multivariate analysis", journal of the royal statistical society, series B, 37,129-149.

6) Blackwell1, James Honaker1 and Gary, (2017), "A Unified Approach to Measurement Error and Missing Data" Journal of American Statistical and Applications, Vol. 46(3), 303-341.

7) Cham, L.S. and Dunn, O.J (1972), "The treatment of missing values in discriminate analysis-I: The sampling experiment" Journal of American Statistical Association, Vol.67, pp.473-477.

8) Cham, L.S. and Dunn, O.J (1974), "A note on the asymptotic aspects of the treatment missing values in discriminate analysis-I: The sampling experiment" Journal of American Statistical *Association*, Vol.69, pp.672-673.

9) Dagenais, M. G (1973)"The use of incomplete observations in multiple regression analysis, A generalized least squares approach, Journal of Econometrics,1 PP, 317-328.

10) El-Dereny, M., &Rashwan, N. (2011), "Solving multicollinearity problem using ridge regression models", *International Journal of Contemporary Mathematical Sciences*. N.I. 4, 447-456.

11) Gold Stein, M., & Smith, A.F.M. (1974) Ridge-type estimators for regression analysis. *Journal of the Royal Statistical Society: Series B (Methodological)*, 36, 284-291.

12) G. Can J (2015), "Stat Variable Selection and Inference Procedures for Marginal Analysis of Longitudinal Data with Missing Observations and Covariate Measurement Error", 43(4):498-518.

13) Hartley, H.O and Hocking, R.R (1971), "The Analysis of incomplete data", Biometric, 27, 783-823.

ISSN:0971-1260
Vol-22- Issue-14-December-2019

14) Hoerl, A.E., & Kennard, (1970a) Ridge regression: Biased estimation for non-orthogonal problems R.W. *Technometric*, 12, 55-67.

15) Hoerl, A.E., & Kennard, (1970b). Ridge regression: Applications to non-orthogonal problems. E*conometric*, R.W. 12, 69-82.

16) Lindley, D.V.(1947), "Regression lines and the linear functional relationship", Journal of the Royal Statistical Society, Series B, Vol.9, PP.218-244.

17) Lankipalli, K.N., Reddy and Ravi .K (1981), "Deficient Data Analysis: Missing Observations", proceedings of the 18th annual Conference of the Indian Econometric society held at Pune.

18) Lankipalli, K.N., and V.K. Reddy, and V.B. Naidu, (1985) "Adjusted linear estimation: Non-Normal case", proceedings of the 23th Conference of the Indian Econometrics society.

19) Lankipalli, V.B. Naidu, K.N. and BalaSiddhamuni .P (1992), "Adjusted linear estimation: An application to form management Data", ArthaVijnana, vol.34, 41-56.

20) Maher EI-Masri, Susan M Fox-Wasylyshyn (2005), "Missing data: An introductory conceptual overview for the no-voice research", Can J Nurse Health, Res. 37(4); pp156-171.

21) Mohd Mustafa AI Bakri Abdullah,(2014)"Filling missing data using interpolation methods: Study on the effect of fitting distribution", Key engineering materials, Vol.594-595, PP.889-895.

22) Naidu, V.B. (1985) "Some problems in variables in linear model", Ph.D. Thesis, S.V. University, Tirupati, India.

23) Orchard, T and Woodbury, M.A. (1972), "A missing observation principle: Theory and Practice", proceedings of the 6th Berkley Symposium, Vol. I, 697-715.

24) Press, S.J. and Scott, A.J (1976), "Missing variables in Bayesian Regression=II", Journal of the American Statistical Association, 71,366-369.

25) Paul D Allison (2012), "Asymmetric Fixed-effects Models for panel Data", Sage Journals, SAS institute 2nd edition.

ISSN:0971-1260
Vol-22- Issue-14-December-2019

26) Rubin, D.B. (1974), “Characterizing the estimation of parameters in incomplete data problem”, Journal of American Statistical Association Vol. 69, pp.467-474.

27) Swami, P.A.V.B. and Mehta, J.S (1975), “On Bayesian Estimation of seeming unrelated regressions, when sum observations are missing”, Journal of Econometrics, 3, 157-170.

28) V.B. Naidu, and V.K. Reddy, (2006) “Estimation of parameters in the presence of Measurement Errors and Missing Data: A New Estimator”, Asian-African Journal of Economics and Econometrics, Vol. 6, No 2, 219-225.

29) Woodbury, M.A. and Saber (1960), “Factor Analysis with missing data”, Annals of New York, Academy of Sciences, 128,764.

 