AD-A114 533   WISCONSIN UNIV-MADISON MATHEMATICS RESEARCH CENTER   F/G 12/1
ON JOINTLY ESTIMATING PARAMETERS AND MISSING DATA BY MAXIMIZING--ETC(U)
FEB 82   R J LITTLE, D B RUBIN   DAAG29-80-C-0041
UNCLASSIFIED   MRC-TSR-2326   NL

1 OF 1
AD
A114533

END
DATE
FILMED
6-82
DTIC

(2)

AD A114533

MRC Technical Summary Report #2326

ON JOINTLY ESTIMATING PARAMETERS AND MISSING DATA BY MAXIMIZING THE COMPLETE-DATA LIKELIHOOD

Roderick J. A. Little
and
Donald B. Rubin

Mathematics Research Center
University of Wisconsin—Madison
610 Walnut Street
Madison, Wisconsin 53706

February 1982

(Received November 9, 1981)

Sponsored by

U. S. Army Research Office
P. O. Box 12211
Research Triangle Park
North Carolina 27709

UNIVERSITY OF WISCONSIN - MADISON
MATHEMATICS RESEARCH CENTER

ON JOINTLY ESTIMATING PARAMETERS AND MISSING DATA BY MAXIMIZING THE COMPLETE-DATA LIKELIHOOD

Roderick J. A. Little and Donald B. Rubin

Technical Summary Report #2326

February 1982

ABSTRACT

One approach to handling incomplete data occasionally encountered in the literature is to treat the missing data as parameters and to maximize the complete data likelihood over missing data and parameters. This paper points out that although this approach can be useful in particular problems, it is not a generally reliable approach to the analysis of incomplete data. In particular, it does not share the optimal properties of maximum likelihood estimation, except under the trivial asymptotics in which the proportion of missing data goes to zero as the sample size increases.

AMS (MOS) Subject Classification: 62A10, 62F10, 62H12

Key Words: Incomplete data, maximum likelihood estimation

Work Unit Number 4 - Statistics and Probability

Sponsored by the United States Army under Contract No. DAAG29-80-C-0041.

ON JOINTLY ESTIMATING PARAMETERS AND MISSING DATA
BY MAXIMIZING THE COMPLETE-DATA LIKELIHOOD

Roderick J. A. Little and Donald B. Rubin

1. Introduction

In the standard formulation of maximum likelihood theory for complete data, the data z are assumed to have a distribution with density $f(z|\theta)$ indexed by an unknown parameter θ. Having observed data values $z = \tilde{z}$, the likelihood of θ is the density of the observed data regarded as a function of θ, that is

$$L(\theta|\tilde{z}) = f(\tilde{z}|\theta) \quad \text{for all} \quad \theta \ . \qquad (1)$$

The maximum likelihood estimate $\hat{\theta}$ of θ is obtained by maximizing (1) with respect to θ. We use the term complete data likelihood to refer to the expression (1).

Now suppose that some of the values in z are not observed. Let z_m denote the missing components and z_p the observed (present) components where $\tilde{z}_p$ is the observed value of z_p. It is not uncommon in the literature on incomplete data to see the suggestion that estimates of θ can be found by treating the missing values z_m as parameters and maximizing the complete data likelihood with respect to θ and z_m. In symbols, this corresponds to maximizing the function

$$L_1(\theta, z_m|\tilde{z}_p) = f(z_m, \tilde{z}_p|\theta) \qquad (2)$$

with respect to (θ, z_m). The classic example of this approach is in the analysis of missing plots in analysis of variance where missing outcomes z_m are treated as parameters and then filled in to allow computationally efficient methods to be used for analysis (Anderson, 1946; Bartlett, 1937;

Sponsored by the United States Army under Contract No. DAAG29-80-C-0041.

Rubin, 1972). More recently, DeGroot and Goel (1980) propose this approach as one possibility for the analysis of a mixed up bivariate normal sample, where the missing data are the indices that allow the values of the two variables to be paired, and a priori all pairings are assumed equally likely. Press and Scott (1976) present a Bayesian analysis of an incomplete multivariate normal sample which is formally equivalent to maximizing (2). They maximize the joint posterior distribution of θ and z_m, after specifying a flat prior distribution the parameter θ.

Although the literature on missing plot analysis explicitly recognizes the problems resulting from the suggested procedure, the more recent literature can be read as implying that maximizing (2) over missing data and parameters is just as principled as standard maximum likelihood estimation from the complete data. Our purpose is simply to point out the joint maximization over missing data and parameters is <u>not</u> a maximum likelihood procedure in any useful sense of the term. It does not in general enjoy the optimal large sample properties of maximum likelihood estimation, except using the trivial asymptotics in which the fraction of the data which are missing goes to zero as the sample size increases.

From the likelihood perspective, missing data z_m differ fundamentally from parameters θ in that they are random variables with an a priori specified probability distribution. The correct likelihood is obtained by integrating the missing data z_m out of the complete data likelihood (1), that is, the correct likelihood is

$$L_2(\theta|\tilde{z}_p) = \int f(z_m, \tilde{z}_p|\theta)dz_m, \quad \text{for all } \theta . \tag{3}$$

This formulation implicitly assumes that the missing data are missing at random (Rubin, 1976). In particular, the probability that a value is missing does not depend on the missing data z_m, although it may depend on values z_p

which are observed. If the missing data are not missing at random, then the model formulation needs to include a distribution for the set of variables indicating whether values are observed or missing. For details, see Rubin (1976).

Assuming the missing data are missing at random, L_2 given by (3) is equal to the probability density of the observed data z_p regarded as a function of the unknown parameter, that is, of quantities not having a probability distribution. Hence L_2 and not L_1 is the true likelihood of θ given incomplete data $\tilde{z}_p$. In the next section we compare parameter estimates of θ found by maximizing L_1 with maximum likelihood estimates found by maximizing L_2 for some simple problems.

2. Examples

Example 1. Univariate Normal Sample

Suppose that z consists of N observations from a Normal distribution with mean μ and variance σ^2, z_p consists of n observations which are observed and z_m represents $N-n$ missing observations which are assumed missing at random. Let $\bar{z}$ and s_z^2 denote the sample mean and sample variance (with denominator n) of the n observed values. Then $\theta = (\mu, \sigma^2)$, and maximizing L_2 leads to maximum likelihood estimates $\mu = \bar{z}$, $\sigma^2 = s_z^2$. In contrast, maximizing L_1 with respect to θ and z_m yields a common estimate $\bar{z}$ for all components of z_m, and estimates $\mu = \bar{z}$, $\sigma^2 = s_z^2(n/N)$. Thus the maximum likelihood estimate of the mean is obtained, but the maximum likelihood estimate of the variance is multiplied by the fraction of observed data. When the fraction of missing data is substantial (for example, $n/N = 0.5$), the estimated variance σ^2 is badly biased, and this bias does not vanish as $N \to \infty$ unless $n/N \to 0$; more relevant asymptotics would fix n/N as the sample size increases.

Example 2. Missing Plot Analysis of Variance

Suppose we add to the previous example a set of covariates x which is observed for all N observations. We assume that the value of z for observation i with covariate values x_i is Normal with mean $\beta_0 + \beta^T x_i$ and variance σ^2. The estimates of β_0 and β obtained by maximizing L_1 are the maximum likelihood estimates, obtained by least squares regression with the n observed data points. However, as in Example 1, the estimate of variance is the maximum likelihood estimate multiplied by the proportion of observed values.

These results provide one justification for the analysis of missing plots in analysis of variance designs mentioned in section 1: jointly estimating the values of the outcome variable for the missing plots and the parameters leads to maximum likelihood estimates of the effects β. However an adjustment is needed to the resulting estimate of the residual variance σ^2, as the literature on missing plot analysis explicitly recognizes.

Example 3. An Exponential Sample

In the first two examples estimation based on maximizing L_1 at least yields reasonable estimates of location, even though estimates of the scale parameter need adjustment. However in other examples, estimates of location can also be biased. For example, consider a censored sample from an exponential distribution with mean μ, where z_p represents the n observed values which lie below a known censoring point c, and z_m represents the $N-n$ values beyond c which are censored. The maximum likelihood estimate of μ is $\hat{\mu} = \bar{z} + (N-n)c/n$. Maximization of L_1 leads to estimating censored values of z at the censoring point c, and estimating μ by $(n/N)\hat{\mu}$. Thus in this case the estimate of the mean is inconsistent unless the proportion of missing values tends to zero as the sample size increases.

Example 4. A Bivariate Normal Sample with Missing Predictor Variables.

Biased estimates of location parameters can also occur in problems involving the normal distribution. For example, suppose that $z_i = (x_i, y_i)$ $i = 1,\ldots,N$ are N observations from a bivariate normal distribution with mean (μ_x, μ_y), variances σ_x^2 and σ_y^2, and correlation ρ, where y_i is observed for all N observations, and $x_1,\ldots,x_n$ are observed but $x_{n+1},\ldots,x_N$ are missing at random. Suppose that interest is focussed on the regression coefficient of y_i on x_i, $\beta_{y.x} = \rho\sigma_y/\sigma_x = \beta_{x.y}\,\sigma_y^2/\sigma_x^2$. The maximum likelihood estimate of $\beta_{y.x}$ is

$$\hat{\beta}_{y.x} = \hat{\beta}_{x.y}\hat{\sigma}_y^2/\hat{\sigma}_x^2 \quad ,$$

where $\hat{\beta}_{x.y} = \sum_{i=1}^{n} (x_i-\bar{x})y_i / \sum_{i=1}^{n} (x_i-\bar{x})^2$, $\bar{x} = \sum_{i=1}^{n} x_i$; $\hat{\sigma}_y^2 = N^{-1}\sum_{i=1}^{N} (y_i-\bar{y})^2$, $\bar{y} = N^{-1}\sum_{i=1}^{N} y_i$; and $\hat{\sigma}_x^2 = \hat{\beta}_{x.y}^2\hat{\sigma}_y^2 + n^{-1}\sum_{i=1}^{n} (x_i-\hat{\beta}_{x.y}y_i)^2$.

Maximization of (2) over parameters and data yields for estimated $\beta_{y.x}$,

$$\hat{\hat{\beta}}_{y.x} = \hat{\beta}_{y.x}\hat{\sigma}_x^2/\hat{\hat{\sigma}}_x^2 \quad ,$$

where $\hat{\hat{\sigma}}_x^2 = \hat{\beta}_{x.y}^2\hat{\sigma}_y^2 + N^{-1}\sum_{i=1}^{n} (x_i-\hat{\beta}_{x.y}y_i)^2$. The estimate $\hat{\hat{\beta}}_{y.x}$ can be badly biased, and again this bias persists as $N \to \infty$ unless the fraction of missing observations tends to zero.

This example is a special case of the problem considered by Press and Scott (1976). They observe that for the general problem they considered their estimates based on maximizing L_2 are consistent only if the fraction of missing observations tends to zero. The correct maximum likelihood approach, as discussed by Trawinski and Bargman (1964), Hartley and Hocking (1971), Orchard and Woodbury (1972), Beale and Little (1975) and Dempster, Laird and Rubin (1977) leads to estimates which are consistent as the sample size increases with the fraction of missing data held constant.

3. Missing Values as Parameters

Both maximum likelihood and the maximization of L_1 over parameters and missing data assumes the existence of a model that specifies a distribution for the observed and missing values of z. Occasionally it is possible that situations will arise when it may be desirable to avoid specifying a distribution for the missing values and to treat them as genuine unknown parameters. Hartley and Hocking (1971, section 4 and 5) discuss the regression of y_i on x_i, where the values x_i correspond to fixed points in an experimental design, y_i is observed for all units i and components of x_i are missing for some units. Writing x_p and x_m for the present and missing values of x, respectively, Hartley and Hocking (1971) suggest drawing inferences by maximizing the complete data likelihood based on the conditional distribution of y given x

$$L_3(\theta, \tilde{x}_m | \tilde{y}, \tilde{x}_p) = f(\tilde{y} | \tilde{x}_m, \tilde{x}_p; \theta) \tag{4}$$

with respect to x_m and the parameters θ. Hartley and Hocking discuss analyses where values of x_m are unconstrained or are constrained to be any of k alternatives. We believe that in most practical situations it is more natural to include a distribution for the missing values in the model (Rubin, 1971). From a strict likelihood perspective, however, there is no reason in principle to reject inferences based on (4). The question of whether x_m should be treated as fixed or integrated out of the likelihood (as in (2)) relates to the more general issue of statistical inference in the presence of nuisance parameters, which lies outside the scope of this note.

REFERENCES

Anderson, R. L. (1946). Missing-plot techniques. Biometrics 2, 41-47.

Bartlett, M. S. (1937). Some examples of statistical methods of research in agriculture and applied botany. J. Roy. Statist. Soc. Ser B 4, 137-170.

Beale, E. M. L. and Little, R. J. A. (1975). Missing values in multivariate analysis. J. Roy. Statist. Soc. Ser B 37:127-146.

Dempster, A. P., Laird, N. and Rubin, D. B. (1977). Maximum Likelihood from Incomplete Data via the EM algorithm. J. Roy. Statist. Soc. - B, 39, 1-38.

DeGroot, M. H. and Goel, K. (1980). Estimation of the Correlation Coefficient from a Broken Random Sample. Ann. Statist., 8, 264-278.

Hartley, H. O. and Hocking, R. R. (1971). The analysis of incomplete data. Biometrics 27, 783-808. (With discussion).

Rubin, D. B. (1971). Discussion of "The analysis on incomplete data", by H. O. Hartley and R. R. Hocking. Biometrics 27, 817-818.

__________ (1972). A non-iterative algorithm for least squares estimation of missing values in any analysis of variance design. J. Roy. Statist. Soc. Ser. C 21, 136-141.

__________ (1976). Inference and missing data. Biometrika 63, 581-592.

Orchard, T. and Woodbury, M. A. (1972). A missing information principle: theory and applications. Proc. 6th Berkeley Symposium on Math. Statist. and Prob. 1, 697-715.

Press, S. J. and Scott, A. J. (1976). Missing Variables in Bayesian Regression, II. J. Am. Statist. Assoc. 71, 366-369.

Trawinski, I. M. and Bargman, R. E. (1964). Maximum likelihood estimates with incomplete data. Ann. Math. Statist. 35, 647-657.

RJAL/DBR/jvs

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

REPORT DOCUMENTATION PAGE		READ INSTRUCTIONS BEFORE COMPLETING FORM
1. REPORT NUMBER #2326	2. GOVT ACCESSION NO. AD-A114 533	3. RECIPIENT'S CATALOG NUMBER
4. TITLE (and Subtitle) On Jointly Estimating Parameters and Missing Data by Maximizing the Complete-Data Likelihood		5. TYPE OF REPORT & PERIOD COVERED Summary Report - no specific reporting period
		6. PERFORMING ORG. REPORT NUMBER
7. AUTHOR(s) Roderick J. A. Little and Donald B. Rubin		8. CONTRACT OR GRANT NUMBER(s) DAAG29-80-C-0041
9. PERFORMING ORGANIZATION NAME AND ADDRESS Mathematics Research Center, University of 610 Walnut Street Wisconsin Madison, Wisconsin 53706		10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS Work Unit Number 4 - Statistics and Probability
11. CONTROLLING OFFICE NAME AND ADDRESS U. S. Army Research Office P.O. Box 12211 Research Triangle Park, North Carolina 27709		12. REPORT DATE February 1982
		13. NUMBER OF PAGES 7
14. MONITORING AGENCY NAME & ADDRESS (if different from Controlling Office)		15. SECURITY CLASS. (of this report) UNCLASSIFIED
		15a. DECLASSIFICATION DOWNGRADING SCHEDULE

16. DISTRIBUTION STATEMENT (of this Report)

Approved for public release; distribution unlimited.

17. DISTRIBUTION STATEMENT (of the abstract entered in Block 20, if different from Report)

18. SUPPLEMENTARY NOTES

19. KEY WORDS (Continue on reverse side if necessary and identify by block number)

Incomplete data, maximum likelihood estimation

20. ABSTRACT (Continue on reverse side if necessary and identify by block number)

One approach to handling incomplete data occasionally encountered in the literature is to treat the missing data as parameters and to maximize the complete data likelihood over missing data and parameters. This paper points out that although this approach can be useful in particular problems, it is not a generally reliable approach to the analysis of incomplete data. In particular, it does not share the optimal properties of maximum likelihood estimation, except under the trivial asymptotics in which the proportion of missing data goes to zero as the sample size increases.

DD FORM 1 JAN 73 1473 EDITION OF 1 NOV 65 IS OBSOLETE

UNCLASSIFIED

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

ME

8