ESTIMATING NONLINEAR FUNCTIONAL RELATIONS BY MAXIMUM LIKELIHOOD AND LEAST SQUARES

by

J.W. Sawyer, Jr. and K.L.Q. Read

Department of Biostatistics
University of North Carolina at Chapel Hill

Institute of Statistics Mimeo Series No. 1374

January 1982

Estimating Nonlinear Functional Relations by Maximum Likelihood and Least Squares

J. W. Sawyer and K.L.Q. Read

Institute of Statistics Mimeo Series No. 1374

Errata

Page 1: eq. (1.1) should read

$$x_{i,p+1} = f(x_{i1}, \ldots, x_{ip}, \underset{\sim}{\alpha}), \quad i = 1, 2, \ldots, t$$

Line 4 from bottom: asterisk V as: V^*.

insert footnote at bottom of page:

* If V is taken to refer to arbitrarily chosen vectors $\underset{\sim}{X}_{1i_1}, \ldots, \underset{\sim}{X}_{ti_t}$, concatenated as $\underset{\sim}{X}' = (\underset{\sim}{X}'_{1i_1}, \ldots, \underset{\sim}{X}'_{ti_t})$, the m_i's need not be equal.

Page 2: line 2: V, not Σ.

para 2 line 6: insert after ... of the data: (that is, one

para 2 line 8: parameters). (insert right parenthesis)

Page 4: 6 lines above eq. (2.1): the pr + r functions

eq. (2.1): $\alpha_\nu = g_\nu(\underset{\sim}{x}_{i_1}, \ldots, \underset{\sim}{x}_{i_r}), \quad \nu = 1, 2, \ldots, r$

Page 6: 2 lines below eq. (2.5): any subset $i_1, \ldots, i_r$ of the ...

Page 8: 2nd eq: $$S_N(\underset{\sim}{x}_1, \ldots, \underset{\sim}{x}_t) = \sum_{i=1}^{t} m_i(\overline{\underset{\sim}{X}}_i - \underset{\sim}{x}_i)' \hat{\Sigma}^{-1} (\overline{\underset{\sim}{X}}_i - \underset{\sim}{x}_i)$$

Page 12: 2nd eq: $$\frac{\partial \ell_N}{\partial \sigma_{ij}} = -N\sigma^{ij} + \frac{1}{2} + \mathrm{tr}\left[\underset{\sim}{\Sigma}^{-1}\left[\underset{\sim}{E}_{ij} + \underset{\sim}{E}_{ji}\right]\underset{\sim}{\Sigma}^{-1}\underset{\sim}{B}\right]$$

Estimating Nonlinear Functional Relations by Maximum Likelihood and Least Squares

by

J. W. Sawyer, Jr.[1]

K. L. Q. Read[2]

[1] Department of Biostatistics, School of Hygiene and Public Health, The Johns Hopkins University, Baltimore, Maryland 21205.

[2] Department of Mathematical Statistics and Operations Research, University of Exeter, Exeter EX4 4PU.

SUMMARY

It is assumed that $p + 1$ unknown variables at t points are such that the $(p + 1)^{st}$ variable is given by a nonlinear function, f, of the other p variables and a vector α of structural parameters. Replicate observations on the $p + 1$ variables for each of the t points are assumed to be mutually independent and multivariate normal, with common covariance matrix Σ. For known Σ, obtaining the maximum likelihood estimate $\hat{\alpha}$ of α is equivalent to minimizing a weighted sum of squares with respect to α and pt nuisance parameters. If Σ is replaced by a consistent estimate $\hat{\Sigma}$ in the weighted sum of squares, the weighted least squares estimate $\tilde{\alpha}$ is no longer equal to $\hat{\alpha}$. However, for N total observations, $\sqrt{N}(\hat{\alpha} - \alpha)$ and $\sqrt{N}\,(\tilde{\alpha} - \alpha)$ have the same asymptotic distribution. This result implies that, even if Σ must be estimated, a close relationship exists between maximum likelihood and the generalized least squares procedure which Dolby (1972) has already applied to nonlinear functional relations with Σ known.

KEYWORDS: Nonlinear functional relation, maximum likelihood, weighted least squares, generalized least squares

1. INTRODUCTION

For $(p + 1)$-dimensional vectors $\underset{\sim}{x}_i$, $i = 1, 2, \ldots t$, for an r-dimensional vector $\underset{\sim}{\alpha}$, $r \leq t$, and for known f, let

$$x_{i,p+1} = f\left(x_i, \ldots x_{ip}, \underset{\sim}{\alpha}\right), \quad i = 1, 2, \ldots t \qquad (1.1)$$

hold for some unknown nonrandom vectors $\underset{\sim}{x}_1^o, \ldots \underset{\sim}{x}_t^o$, and $\underset{\sim}{\alpha}^o$. Let observations

$$\underset{\sim}{X}_{ik} = \underset{\sim}{x}_i^o + \underset{\sim}{\epsilon}_{ik}, \quad k = 1, 2, \ldots m_i; i = 1, \ldots t, \qquad (1.2)$$

with $E(\underset{\sim}{\epsilon}_{ik}) = \underset{\sim}{0}$, exist on each $\underset{\sim}{x}_i^o$. Then the observations (1.2) are said to have a functional relationship. Most treatments of the problem of estimating $\underset{\sim}{\alpha}_o$ have assumed that each $\underset{\sim}{X}_{ik}$ is multivariate normal.

If t is fixed while each m_i becomes large, then maximum likelihood is an appropriate method of estimation. Dolby and Freeman (1975) discuss maximum likelihood estimation for linear or nonlinear f, with all $m_i = m > 1$. They allow for correlation between $\underset{\sim}{X}_{ik}$ and $\underset{\sim}{X}_{jk}$ for $i \neq j$, but require that $\underset{\sim}{X}_{ik}$ and $\underset{\sim}{X}_{jh}$ be independent for $k \neq h$, even if $i = j$. Let $\underset{\sim}{X}'_k = (\underset{\sim}{X}'_{1k}, \ldots \underset{\sim}{X}'_{tk})$, and let all $\underset{\sim}{X}_k$ have variance-covariance matrix $\underset{\sim}{V}$.* Then another method of estimating $\underset{\sim}{\alpha}_o$ is to calculate $\tilde{\underset{\sim}{\alpha}}$ minimizing the weighted sum of squares

$$\sum_{k=1}^{m} \left(\underset{\sim}{X}_k - \tilde{\underset{\sim}{x}} \right)' \underset{\sim}{V}^{-1} \left(\underset{\sim}{X}_k - \tilde{\underset{\sim}{x}} \right), \qquad (1.3)$$

---

* If $\underset{\sim}{V}$ is taken to refer to arbitrarily chosen vectors $\underset{\sim}{X}_{1i_1}, \ldots, \underset{\sim}{X}_{ti_t}$, concatenated as $\underset{\sim}{X}' = (\underset{\sim}{X}'_{1i_1}, \ldots, \underset{\sim}{X}'_{ti_t})$, the m_i's need not be equal.

where $\tilde{\underset{\sim}{x}}' = (\tilde{\underset{\sim}{x}}_1', \ldots \tilde{\underset{\sim}{x}}_t')$ satisfies the equations (1.1). For known $\underset{\sim}{\Sigma}$ and multivariate normal observations, $\tilde{\underset{\sim}{\alpha}}$ is clearly the maximum likelihood estimate. The need to estimate the pt nuisance parameters $x_{11}^o, \ldots x_{1p}^o, x_{21}^o, \ldots x_{tp}^o$ can be eliminated by a modification of (1.3) which involves first order Taylor series approximations of the constraints (1.1). For p = 1 and known $\underset{\sim}{V}$, Dolby (1972) demonstrates that this approach, known as generalized least squares, is equivalent to maximum likelihood, except for the error arising from the first order linear approximation to a nonlinear f. For any p and for unknown $\underset{\sim}{V}$, Dolby and Freeman (1975) establish that the maximum likelihood and generalized least square estimates are equivalent if f is linear.

The equivalence to within an approximation of generalized least squares and maximum likelihood for nonlinear f and known $\underset{\sim}{V}$, which Dolby (1972) establishes for p = 1, can easily be extended to the case of arbitrary p by a similar development. Now suppose that a consistent estimate $\hat{\underset{\sim}{V}}$ of $\underset{\sim}{V}$ replaces $\underset{\sim}{V}$ in (1.3), and further suppose that $\hat{\underset{\sim}{V}}$ is a fixed, never-updated function of the data (that is, one which does not involve iterated estimates of $\underset{\sim}{\alpha}_o$ or the nuisance parameters). The estimate $\tilde{\underset{\sim}{\alpha}}$ obtained by minimizing (1.1) will still be equivalent to a generalized least squares estimate based on $\hat{\underset{\sim}{V}}$ instead of $\underset{\sim}{V}$, to within a first order approximation. This can be established by viewing the negative of (1.3) as a "likelihood" to be maximized, and employing the same arguments as Dolby applies to the likelihood when $\underset{\sim}{V}$ is known.

The primary purpose of this paper is to demonstrate that, even if $\underset{\sim}{V}$ must be estimated, minimization of (1.3) is still as precise a procedure as maximum likelihood, in the sense that $\tilde{\underset{\sim}{\alpha}}$ and the maximum likelihood estimate have the same asymptotic distribution. In the light of the discussion above, this means that any asymptotic inefficiency involved in using generalized least squares as opposed to maximum likelihood for nonlinear f will be due only to the approximation inherent in the generalized least squares estimate.

For normal data the equality of the asymptotic distributions of a weighted least squares estimate $\tilde{\underset{\sim}{\alpha}}$, employing random weights, and the maximum likelihood estimate, will be established in Section 4. Some preliminary results necessary for Section 4 will be presented in Section 2, along with some results necessary to Section 3, which establishes the consistency of $\tilde{\underset{\sim}{\alpha}}$ without assuming that the data are normal.

2. PRELIMINARIES

The formal model to be treated is as follows: Let (1.1) and (1.2) hold, and further require that $\underset{\sim}{X}_{ik}$ and $\underset{\sim}{X}_{jh}$ be independent unless $i = j$ and $k = h$. Let all $\underset{\sim}{X}_{ik}$ have non-singular covariance matrix $\underset{\sim}{\Sigma}$. Thus $\underset{\sim}{V}$ becomes block diagonal, with all diagonal blocks equal to $\underset{\sim}{\Sigma}$. (This model is intuitively appealing in that it allows sums of squares and cross products to be pooled in order to estimate $\underset{\sim}{\Sigma}$. With minor modifications, the arguments which follow will work just as well for the model with general $\underset{\sim}{V}$ treated

in Dolby and Freeman (1975), or for a model which allows a separate matrix $\underset{\sim}{\Sigma}_i$ for each set of replications $\underset{\sim}{X}_{i1}, \ldots \underset{\sim}{X}_{im_i}$, $i = 1, 2, \ldots t$). Let all m_i increase in fixed proportions, that is, for $N = \Sigma m_i$, let $m_i = \pi_i N$ for fixed π_i.

Let f possess first partials with respect to all its arguments which are continuous in a neighborhood of the true values. For every choice $i_1, \ldots i_r$ of the integers from 1 to t, let the Jacobian of the matrix with elements

$$\frac{\partial x_{i_\mu, p+1}}{\partial \alpha_\nu}, \quad \mu, \nu = 1, 2, \ldots r$$

be different from 0. It will follow that, in a neighborhood of the true values, the pt + r functions

$$x_{i_\mu, p+1} = f\left(x_{i_\mu 1}, \ldots x_{i_\mu p}, \underset{\sim}{\beta}\right) \quad \mu = 1, 2, \ldots r$$

and

$$x_{i_\mu j} = x_{i_\mu j} \qquad \mu = 1, 2, \ldots r, \; j = 1, 2, \ldots p$$

will have an inverse possessing continuous first partials, and in particular,

$$\beta_\nu = g_\nu\left(\underset{\sim}{x}_{i_1}, \ldots \underset{\sim}{x}_{i_r}\right), \quad \nu = 1, 2, \ldots r \qquad (2.1)$$

for some vector function $\underset{\sim}{g}$ which possesses continuous first partials. It follows that, within some neighborhood of the true values any vectors $\underset{\sim}{x}_1, \ldots \underset{\sim}{x}_t$ satisfying (1.1) are such that the corresponding $\underset{\sim}{\alpha}$ may be obtained by anyone of $\binom{t}{r}$ possible functions $\underset{\sim}{g}$.

This result can be used to establish that if some estimates $\tilde{\underset{\sim}{x}}_1, \ldots \tilde{\underset{\sim}{x}}_t$ of the true population means satisfy (1.1), and if the differences $\tilde{\underset{\sim}{x}}_1 - \underset{\sim}{x}_1^o, \ldots \tilde{\underset{\sim}{x}}_t - \underset{\sim}{x}_t^o$ are $O_p(N^{-\frac{1}{2}})$, then $\tilde{\underset{\sim}{\alpha}} - \underset{\sim}{\alpha}_o$ is $O_p(N^{-\frac{1}{2}})$ also.

This result in turn will be necessary in order to apply a result of Weiss (1971). Let a data vector $\underset{\sim}{Z}_n$ have log-likelihood $\ell(\underset{\sim}{Z}_n, \underset{\sim}{\theta}^o)$ for a q-dimensional vector $\underset{\sim}{\theta}_o$. Let $\tilde{\underset{\sim}{\theta}}_n$ be an estimate of $\underset{\sim}{\theta}^o$ satisfying

$$\lim_{n \to \infty} P\{\sqrt{n}|\tilde{\theta}_{in} - \theta_i^o| \geq L(n)\} \to 0 \qquad (2.2)$$

for $i = 1, 2, \ldots q$ and all non-random sequences $\{L(n)\}$ with $L(n) \to \infty$. (Condition (2.2) holds if and only if $\tilde{\theta}_{in} - \theta_i^o$ is $O_p(n^{-\frac{1}{2}})$). Let $A_n(\tilde{\underset{\sim}{\theta}}_n)$ be the vector $\partial\ell/\partial\underset{\sim}{\theta}$ evaluated at $\tilde{\underset{\sim}{\theta}}_n$, and let $\underset{\sim}{I}_n(\tilde{\underset{\sim}{\theta}}_n)$ be the $q \times q$ information matrix evaluated at $\tilde{\underset{\sim}{\theta}}_n$. Under mild regularity conditions, Weiss demonstrates that

$$\underset{\sim}{\theta}_n^* = \tilde{\underset{\sim}{\theta}}_n + [\underset{\sim}{I}(\tilde{\underset{\sim}{\theta}}_n)]^{-1} \underset{\sim}{A}(\tilde{\underset{\sim}{\theta}}_n) \qquad (2.3)$$

is such that $\sqrt{n}(\underset{\sim}{\theta}_n^* - \underset{\sim}{\theta}^o)$ has the same asymptotic distribution as $\sqrt{n}(\hat{\underset{\sim}{\theta}}_n - \underset{\sim}{\theta}^o)$, where $\hat{\underset{\sim}{\theta}}_n$ is the maximum likelihood estimate.

Now if n is equated with N for the model described at the beginning of this section, and the data are multivariate normal, $\underset{\sim}{\theta}^o$ will be the vector of dimension $pt + r + (p+1)(p+2)/2$ comprising $\underset{\sim}{\alpha}$, the pt free parameters $x_{11}^o, \ldots x_{1p}^o, x_{21}^o, \ldots x_{tp}^o$, and the free parameters of $\underset{\sim}{\Sigma}$. Further, $\underset{\sim}{I}_N(\underset{\sim}{\theta}^o)/N$ will be continuous and non-random

for all permissible N, and $\underset{\sim}{I}_N(\tilde{\underset{\sim}{\theta}}_N)/N$ will converge to $\underset{\sim}{I}_N(\underset{\sim}{\theta}^o)/N$ in probability. Now (2.3) can be rewritten

$$\sqrt{N}(\underset{\sim}{\theta}^*_N - \underset{\sim}{\theta}^o) - \sqrt{N}(\tilde{\underset{\sim}{\theta}}_N - \underset{\sim}{\theta}^o) = (I_N(\tilde{\underset{\sim}{\theta}}_N)/N)^{-1}(\underset{\sim}{A}_N(\tilde{\underset{\sim}{\theta}}_N)/\sqrt{N}).$$

It follows by Slutsky's theorem and a theorem of Cramér (1946) that if

$$\underset{\sim}{A}_N(\tilde{\underset{\sim}{\theta}}_N)/\sqrt{N} = (\partial \ell/\partial \tilde{\underset{\sim}{\theta}}_N)/\sqrt{N} \qquad (2.4)$$

converges in probability to 0, $\sqrt{N}(\underset{\sim}{\theta}^*_N - \underset{\sim}{\theta}^o)$ and $\sqrt{N}(\tilde{\underset{\sim}{\theta}}_N - \underset{\sim}{\theta}^o)$, and hence $\sqrt{N}(\hat{\underset{\sim}{\theta}}_N - \underset{\sim}{\theta}^o)$ and $\sqrt{N}(\tilde{\underset{\sim}{\theta}}_N - \underset{\sim}{\theta}^o)$, will have the same asymptotic distribution.

The following simple result establishes that $\tilde{\underset{\sim}{\alpha}}$ meets the criteria (2.2) if the corresponding vectors $\tilde{\underset{\sim}{x}}_1, \ldots \tilde{\underset{\sim}{x}}_t$ satisfying (1.1) do. Let $h_1, \ldots h_s$ be non-negative functions of random variables $Y_1, \ldots Y_s$, respectively. Then for nonrandom y

$$P\left(\sum_{i=1}^{s} h_i(Y_i) \geq y\right) \leq \sum_{i=1}^{s} P\left(h_i(Y_i) \geq y/s\right) \qquad (2.5)$$

A simple proof of (2.5) for s = 2 is given in Tucker (1967). Now chose any subset $i_1, \ldots i_r$ at the subscripts 1 to t. Then with probability approaching 1, $\tilde{\underset{\sim}{x}}_{i_1}, \ldots \tilde{\underset{\sim}{x}}_{i_r}$ will lie in the neighborhood for which the associated function g, as defined by (2.1), exists. Thus, with probability approaching 1,

$$\sqrt{N}\left(\tilde{\alpha}_\mu - \alpha^o_\mu\right) = \sum_{\nu=1}^{r} \sqrt{N}\left(\tilde{\underset{\sim}{x}}_{i_\nu} - \underset{\sim}{x}^o_{i_\nu}\right)' \left[\partial g_\mu/\partial \underset{\sim}{x}^*_{i_\nu}\right]$$

where the vector $(\underset{\sim}{x}^{*\prime}_{i_1}, \ldots \underset{\sim}{x}^{*\prime}_{i_r})$ lies on the line segment joining $(\underset{\sim}{x}^{o\prime}_{i_1}, \ldots \underset{\sim}{x}^{o\prime}_{i_r})$ and $(\tilde{\underset{\sim}{x}}'_{i_1}, \ldots \tilde{\underset{\sim}{x}}'_{i_r})$. Since each vector of partials $\partial g_\mu / \partial \underset{\sim}{x}^{*}_{i_\nu}$ is continuous, it will converge in probability to $\partial g_\mu / \partial \underset{\sim}{x}^{o}_{i_\nu}$. Thus, with probability approaching 1, the absolute values of the partials $\partial g / \partial x^{*}_{i_\nu j}$ are bounded by some constant M which does not depend on the data. It follows that, for any nonrandom sequence $\{L(N)\}$ with $L(N) \to \infty$

$$P\left(\sqrt{N}\left|\tilde{\underset{\sim}{\alpha}}_\mu - \underset{\sim}{\alpha}_\mu^{\,o}\right| \geq L(N)\right) \leq P\left(M \sum_{\nu=1}^{r} \sum_{j=1}^{p+1} \sqrt{N}\left|\tilde{x}_{i_\nu j} - x^{o}_{i_\nu j}\right| \geq L(N)\right)$$

with probability approaching 1 as $N \to \infty$. Hence, if the differences $\tilde{\underset{\sim}{x}}_1 - \underset{\sim}{x}^{o}_1, \ldots \tilde{\underset{\sim}{x}}_t - \underset{\sim}{x}^{o}_t$ are $0_p(N^{-\frac{1}{2}})$, we have

$$\lim_{N \to \infty} P\left(\sqrt{N}\left|\tilde{\underset{\sim}{\alpha}}_\mu - \underset{\sim}{\alpha}_\mu^{\,o}\right| \geq L(N)\right) = 0 \qquad (2.6)$$

by virtue of (2.5).

From the results of this section, it follows that, for Section 3, it suffices to show that $\tilde{\underset{\sim}{x}}_1, \ldots \tilde{\underset{\sim}{x}}_t$ minimizing

$$\sum_{i=1}^{t} \sum_{k=1}^{m_i} \left(\underset{\sim}{X}_{ik} - \underset{\sim}{x}_i\right)' \hat{\underset{\sim}{\Sigma}}^{-1} \left(\underset{\sim}{X}_{ik} - \underset{\sim}{x}_i\right) \qquad (2.7)$$

subject to (1.1) are such that $\tilde{\underset{\sim}{x}}_1 - \underset{\sim}{x}_1^{\,o} \ldots \tilde{\underset{\sim}{x}}_t - \underset{\sim}{x}_t^{\,o}$ are $0_p(N^{-\frac{1}{2}})$ for suitably chosen $\hat{\underset{\sim}{\Sigma}}$. In Section 4, it will suffice to show that the functions (2.4) associated with the likelihood of the observations $\underset{\sim}{X}_{ik}$, $i = 1, 2, \ldots t$, $j = 1, 2, \ldots, m_i$, converge to 0 on probability when evaluated at the weighted least squares solutions and $\hat{\underset{\sim}{\Sigma}}$.

3. CONSISTENCY OF THE LEAST SQUARES ESTIMATORS

For normal data, (2.7) is equivalent to a likelihood conditional on an estimated $\hat{\underset{\sim}{\Sigma}}$. Consistency and possibly efficiency of $\tilde{\underset{\sim}{\alpha}}$ might be argued from that viewpoint. Postulating a distribution is not required in order to demonstrate that $\tilde{\underset{\sim}{\alpha}} - \underset{\sim}{\alpha}^{o}$ is $0_p(N^{-\frac{1}{2}})$, however. It is only required that $\hat{\underset{\sim}{\Sigma}} - \underset{\sim}{\Sigma}$ be elementwise $0_p(N^{-\frac{1}{2}})$, that it be positive definite with probability approaching 1, and that its definition not be dependent on the minimization of (2.7).

Now (2.7) is equal to

$$\sum_{i=1}^{t} \sum_{k=1}^{m_i} \left(\underset{\sim}{X}_{ik} - \overline{\underset{\sim}{X}}_i\right)' \hat{\Sigma}^{-1} \left(\underset{\sim}{X}_{ik} - \overline{\underset{\sim}{X}}_i\right) + S_N\left(\underset{\sim}{x}_1, \ldots \underset{\sim}{x}_t\right)$$

where $\overline{\underset{\sim}{X}}_i$ is the mean of the m_i replications and

$$S_N\left(\underset{\sim}{X}_1, \ldots \underset{\sim}{x}_t\right) = \sum_{i=1}^{t} m_i \left(\overline{\underset{\sim}{X}}_i - \underset{\sim}{x}_i\right)' \hat{\Sigma}^{-1} \left(\overline{\underset{\sim}{X}}_i - \underset{\sim}{x}_i\right)$$

Thus it is sufficient to determine $\tilde{\underset{\sim}{x}}_1, \ldots \tilde{\underset{\sim}{x}}_t$ minimizing S_N under the constraints (1.1). Let $\underset{\sim}{\Sigma} = \{\sigma_{jh}\}$ and $\underset{\sim}{\Sigma}^{-1} = \{\sigma^{jh}\}$. Then $S_N(\underset{\sim}{x}_1^{o}, \ldots \underset{\sim}{x}_t^{o})$ is the sum of terms of the form

$$m_i \left(\overline{X}_{ij} - x_{ij}^{o}\right)^2 \hat{\sigma}^{ij}$$

or

$$m_i \left(\overline{X}_{ij} - x_{ij}^{o}\right)\left(\overline{X}_{ih} - x_{ih}^{o}\right)\hat{\sigma}^{jh}$$

for $j \neq h$.

Since each $\hat{\sigma}^{jh}$ is a continuous function of the elements of $\hat{\underset{\sim}{\Sigma}}$,

$\hat{\sigma}^{jh} \to \sigma^{jh}$ in probability for j, h = 1, 2, ... p + 1. Thus, with probability approaching 1, all elements of $\hat{\underset{\sim}{\Sigma}}$ have a nonrandom upper bound M. Since $m_i = \pi_i N$, the mean $\bar{X}_{ij}$ is such that $\bar{X}_{ij} - x_{ij}^{o}$ is $0_p(N^{-\frac{1}{2}})$. It follows that, for any nonrandom sequence $\{L(N)\}$ with $L(N) \to \infty$,

$$P\left(\sqrt{m_i}\,|\bar{X}_{ij} - x_{ij}^{o}| \geq \sqrt{L(N)}\right) = P\left(m_i\left(\bar{X}_{ij} - x_{ij}^{o}\right)^2 \geq L(N)\right) \quad (3.1)$$

approaches 0 as $N \to \infty$. Since

$$2|\bar{X}_{ij} - x_{ij}^{o}|\,|\bar{X}_{ih} - x_{ih}^{o}| \leq |\bar{X}_{ij} - x_{ij}^{o}|^2 + |\bar{X}_{ih} - x_{ih}^{o}|^2 \quad (3.2)$$

a result similar to (3.1) applies to the cross-product terms. In the manner in which we established (2.6), we apply (2.5), (3.1), (3.2), and the boundedness of the elements of $\hat{\underset{\sim}{\Sigma}}^{-1}$ with probability approaching 1 to obtain

$$\lim_{N \to \infty} P\left(S_N\left(\underset{\sim}{x}_1^{o}, \ldots \underset{\sim}{x}_t^{o}\right) \geq L(N)\right) = 0$$

for all nonrandom sequences $\{L(N)\}$ with $L(N) \to \infty$. Since

$$S_N\left(\tilde{\underset{\sim}{x}}_1, \ldots \tilde{\underset{\sim}{x}}\right) \leq S_N\left(\underset{\sim}{x}_1^{o}, \ldots \underset{\sim}{x}_t^{o}\right)$$

it follows that

$$\lim_{N \to \infty} P\left(S_N\left(\tilde{\underset{\sim}{x}}_1, \ldots \tilde{\underset{\sim}{x}}_t\right) \geq L(N)\right) = 0 \quad (3.3)$$

Now let $\underset{\sim}{T}'\underset{\sim}{T} = \hat{\underset{\sim}{\Sigma}}$ be a square root decomposition obtained by some well defined algorithm (such as a Cholesky decomposition), such that the elements of $\underset{\sim}{T}$ are continuous functions of the elements of $\hat{\underset{\sim}{\Sigma}}$.

Let

$$\underset{\sim}{W}_i = \left(T^{-1}\right)' \underset{\sim}{\bar{X}}_i$$

and

$$\underset{\sim}{w}_i = \left(T^{-1}\right)' \underset{\sim}{\tilde{x}}_i$$

Then

$$S_N\left(\underset{\sim}{\tilde{x}}_1, \ldots \underset{\sim}{\tilde{x}}_p\right) = \sum_{i=1}^{t} \sum_{j=1}^{p+1} m_i \left(W_{ij} - w_{ij}\right)^2$$

From (3.3) and the fact that $m_i = \pi_i N$, it follows that, for all nonrandom $\{L(N)\}$ with $L(N) \to \infty$, and for all i and j,

$$\lim_{N \to \infty} P\{\sqrt{N}|W_{ij} - w_{ij}| \geq L(N)\} = 0$$

of necessity.

Now suppose that $\underset{\sim}{\tau}'\underset{\sim}{\tau} = \underset{\sim}{\Sigma}$ is a square root decomposition of $\underset{\sim}{\Sigma}$ obtained by the same algorithm as $\underset{\sim}{T}$. It follows that every element T_{jh} of $\underset{\sim}{T}$ converges in probability to the corresponding element τ_{jh} of $\underset{\sim}{\tau}$. Since

$$\underset{\sim}{\bar{X}}_i - \underset{\sim}{x}_i = \underset{\sim}{T}'\left(\underset{\sim}{W}_i - \underset{\sim}{w}_i\right),$$

it follows that

$$\sqrt{N}\ |\bar{X}_{ij} - \tilde{x}_{ij}| \leq \sum_{h=1}^{p+1} \sqrt{N}\ |W_{ih} - w_{ih}||T_{hj}| .$$

Since the values $|T_{jh}|$ are bounded by some nonrandom constant with probability approaching 1, it follows, using (2.5) that

$$\lim_{N \to \infty} P\left(\sqrt{N}\ |\bar{X}_{ij} - \tilde{x}_{ij}| \geq L(N)\right) = 0 \qquad (3.4)$$

But since

$$\sqrt{N}|\tilde{x}_{ij} - x_{ij}^{o}| \leq \sqrt{N}|\bar{X}_{ij} - x_{ij}^{o}| + \sqrt{N}|\bar{X}_{ij} - \tilde{x}_{ij}|,$$

the assertion

$$\lim_{N \to \infty} P\left(\sqrt{N}|\tilde{x}_{ij} - x_{ij}^{o}| \geq L(N)\right) = 0$$

is established by virtue of (2.5).

4. THE MULTIVARIATE NORMAL CASE

Now let the observations be multivariate normal. A natural estimate for $\underset{\sim}{\Sigma}$ is

$$\hat{\underset{\sim}{\Sigma}} = \frac{1}{N} \sum_{i=1}^{t} \sum_{k=1}^{m_i} \left(\underset{\sim}{X}_{ik} - \bar{\underset{\sim}{X}}_{i}\right)\left(\underset{\sim}{X}_{ik} - \bar{\underset{\sim}{X}}_{i}\right)' .$$

The log-likelihood of the observation is

$$\ell_N = -\left[(p+1)N/2\right] \ell n(2\pi) - (N/2) \ell n|\underset{\sim}{\Sigma}| - \frac{1}{2} \sum_{i=1}^{t} \sum_{k=1}^{m_i} \left(\underset{\sim}{X}_{ik} - \underset{\sim}{x}_{i}\right)' \underset{\sim}{\Sigma}^{-1} \left(\underset{\sim}{X}_{ik} - \underset{\sim}{x}_{i}\right)$$

Now clearly

$$-2 \frac{\partial \ell_N}{\partial \alpha_\mu} = \frac{\partial S_N}{\partial \alpha_\mu} \qquad (4.1)$$

for $\mu = 1, 2, \ldots r$, and

$$-2 \frac{\partial \ell_N}{\partial x_{ij}} = \frac{\partial S_N}{\partial x_{ij}} \qquad (4.2)$$

for $i = 1, 2, \ldots t$, $j = 1, 2, \ldots p$, when ℓ_N is evaluated at $\underset{\sim}{\Sigma} = \hat{\underset{\sim}{\Sigma}}$. When (4.1) and (4.2) are evaluated at $\tilde{\underset{\sim}{\alpha}}$ and the pt estimated nuisance parameters $\tilde{x}_{11}, \ldots \tilde{x}_{1p}, \tilde{x}_{21}, \ldots \tilde{x}_{tp}$, all these

partials are identically 0. Thus the conditions that $\frac{1}{\sqrt{N}} \frac{\partial \ell_N}{\partial \underset{\sim}{\alpha}}$ and $\frac{1}{\sqrt{N}} \frac{\partial \ell_N}{\partial \underset{\sim}{x}_{ij}}$ approach 0 in probability, when evaluated at the least squares estimates and $\hat{\underset{\sim}{\Sigma}}$, are trivially fulfilled. It remains to show that $\frac{1}{\sqrt{N}} \frac{\partial \ell_N}{\partial \sigma_{ij}} \to 0$ in probability when evaluated at these estimates.

Write ℓ_N as

$$\ell_N = -\left[(p+1)N/2\right] \ell n(2\pi) - \frac{1}{2} N \ell n|\underset{\sim}{\Sigma}| - \frac{1}{2} \operatorname{tr}\left[\underset{\sim}{\Sigma}^{-1} \sum_i \sum_k \left(\underset{\sim}{X}_{ik} - \underset{\sim}{x}_i\right)\left(\underset{\sim}{X}_{ik} - \underset{\sim}{x}_i\right)'\right]$$

and let $\sum_i \sum_k \left(\underset{\sim}{X}_{ik} - \underset{\sim}{x}_i\right)\left(\underset{\sim}{X}_{ik} - \underset{\sim}{x}_i\right)' = \underset{\sim}{B}$ for convenience. Let $\underset{\sim}{E}_{ij} = \{e_{kh}\}$ be defined by $e_{kh} = 1$ for $k = i$ and $h = j$ and $e_{kh} = 0$ otherwise. Then for $i < j$,

$$\frac{\partial \ell_N}{\partial \sigma_{ij}} = -N \sigma^{ij} + \frac{1}{2} \operatorname{tr}\left[\underset{\sim}{\Sigma}^{-1}\left[\underset{\sim}{E}_{ij} + \underset{\sim}{E}_{ji}\right]\underset{\sim}{\Sigma}^{-1} \underset{\sim}{B}\right]$$

or

$$\frac{\partial \ell_N}{\partial \sigma_{ij}} = - N \sigma^{ij} + \frac{1}{2} \operatorname{tr}\left[\left(\underset{\sim}{E}_{ij} + \underset{\sim}{E}_{ji}\right)\underset{\sim}{\Sigma}^{-1} \underset{\sim}{B} \underset{\sim}{\Sigma}^{-1}\right].$$

Let

$$\underset{\sim}{\Sigma}^{-1} \underset{\sim}{B} \underset{\sim}{\Sigma}^{-1} = \underset{\sim}{A} = \{a_{ij}\}.$$

Then $\frac{\partial \ell_N}{\partial \sigma_{ij}}$ is given by

$$- N \sigma^{ij} + \frac{1}{2}\left[a_{ij} + a_{ji}\right] = -N \sigma^{ij} + a_{ij} \qquad (4.3)$$

For $i = j$, a similar derviation gives

$$\frac{\partial \ell_N}{\partial \sigma_{ii}} = - \frac{1}{2} N \sigma^{ii} + \frac{1}{2} a_{ii} \qquad (4.4)$$

[The derivations of (4.3) and (4.4) follow closely the usual line for the derivation of the maximum likelihood estimates for the underlying parameters of multivariate normal data. Compare Rao (1973).] For a square matrix $\underset{\sim}{C}$, let diag($\underset{\sim}{C}$) be a diagonal matrix with diagonal elements equal to the diagonal elements of $\underset{\sim}{C}$. Then (4.3) and (4.4) imply

$$\frac{\partial \ell_N}{\partial \underset{\sim}{\Sigma}} = -N\,\underset{\sim}{\Sigma}^{-1} + \frac{1}{2}\,N\,\mathrm{diag}(\Sigma^{-1}) + \underset{\sim}{A} - \frac{1}{2}\,\mathrm{diag}(\underset{\sim}{A}) \qquad (4.5)$$

But $\underset{\sim}{A}$ may be expressed as

$$\underset{\sim}{A} = \underset{\sim}{\Sigma}^{-1}\,[N\,\underset{\sim}{\hat{\Sigma}}]\,\underset{\sim}{\Sigma}^{-1} + \underset{\sim}{\Sigma}^{-1}\left[\sum_i m_i\left(\underset{\sim}{\bar{X}}_i - \underset{\sim}{x}_i\right)\left(\underset{\sim}{\bar{X}}_i - \underset{\sim}{x}_i\right)'\right]\underset{\sim}{\Sigma}^{-1}$$

so that, when (4.5) is evaluated at $\underset{\sim}{\Sigma} = \underset{\sim}{\hat{\Sigma}}$ and the least squares solutions,

$$\frac{\partial \ell_N}{\partial \underset{\sim}{\Sigma}} = \underset{\sim}{\hat{\Sigma}}^{-1}\left[\sum_i m_i\left(\underset{\sim}{\bar{X}}_i - \underset{\sim}{\tilde{x}}_i\right)\left(\underset{\sim}{\bar{X}}_i - \underset{\sim}{\tilde{x}}_i\right)'\right]\underset{\sim}{\hat{\Sigma}}^{-1} - \frac{1}{2}\,\mathrm{diag}\left\{\underset{\sim}{\hat{\Sigma}}^{-1}\sum_i m_i\left(\underset{\sim}{\bar{X}}_i - \underset{\sim}{\tilde{x}}_i\right)\left(\underset{\sim}{\bar{X}}_i - \underset{\sim}{\tilde{x}}_i\right)'\underset{\sim}{\Sigma}^{-1}\right\}$$

Now since the elements of $\underset{\sim}{\hat{\Sigma}}^{-1}$ converge to those $\underset{\sim}{\Sigma}^{-1}$ in probability, it follows that $\frac{1}{\sqrt{N}}\,\frac{\partial \ell_N}{\partial \underset{\sim}{\Sigma}}$ converges to 0 elementwise if

$$\frac{1}{\sqrt{N}}\sum_i m_i\left(\underset{\sim}{\bar{X}}_i - \underset{\sim}{\tilde{x}}_i\right)\left(\underset{\sim}{\bar{X}}_i - \underset{\sim}{\tilde{x}}_i\right)' \xrightarrow{p} 0$$

elementwise. That is, it is sufficient to demonstrate that

$$\sqrt{N}\sum_i \pi_i\left(\bar{X}_{ij} - \tilde{x}_{ij}\right)\left(\bar{X}_{ih} - \tilde{x}_{ih}\right) \xrightarrow{p} 0\ ,\ j, h = 1, 2, \ldots p+1 \qquad (4.6)$$

Now for any prescribed $\epsilon > 0$, choosing $L(N) = \sqrt{\epsilon}\, N^{\frac{1}{4}}$ in (3.4) will yield

$$\lim_{N \to \infty} P\left(\sqrt{N}\left(\bar{X}_{ij} - \tilde{x}_{ij}\right)^2 \geq \epsilon\right) = 0$$

Since $\sqrt{N}\left|\left(\bar{X}_{ij} - \tilde{x}_{ij}\right)\left(\bar{X}_{ih} - \tilde{x}_{ih}\right)\right| \leq \sqrt{N}\left(\bar{X}_{ij} - \tilde{x}_{ij}\right)^2 + \sqrt{N}\left(\bar{X}_{ih} - \tilde{x}_{ih}\right)^2$

(4.6) must be valid. Hence the asymptotic distributions of $\sqrt{N}\left(\tilde{\underset{\sim}{\alpha}} - \underset{\sim}{\alpha}_o\right)$ and $\sqrt{N}\left(\hat{\underset{\sim}{\alpha}} - \underset{\sim}{\alpha}_o\right)$, where $\hat{\underset{\sim}{\alpha}}$ is the ML estimate, are equal.

REFERENCES

Cramér, H. (1946). Mathematical Methods of Statistics. Princeton: Princeton University Press.

Dolby, G. R. (1972). Generalized least squares and maximum likelihood estimation of non-linear functional relationships. J. Roy. Statist. Soc., 34, 393-400.

Dolby, G. R. and Freeman T. G. (1975). Functional relationships having many independent variables and errors with multivariate normal distribution. J. Multivariate Anal. 5, 466-479.

Rao, C. R. (1973). Linear Statistical Inference and Its Applications. New York: John Wiley and Sons.

Tucker, M. G. (1967). A Graduate Course in Probability. New York: Academic Press.

Weiss, L. (1971). Asymptotic properties of maximum likelihood estimators in some nonstandard cases. J. Amer. Statist. Assoc., 66, 345-350.