MAXIMUM LIKELIHOOD AND LEAST SQUARES ESTIMATION

by

Jennifer McDonald DaCosta Goddard

A thesis submitted to the Faculty of Graduate Studies and Research in partial fulfillment of the requirements for the degree of Master of Science

Department of Mathematics
McGill University
Montreal

July, 1978

MAXIMUM LIKELIHOOD AND LEAST SQUARES ESTIMATION

by

JENNIFER McDONALD DaCOSTA GODDARD

RÉSUMÉ

Cette thèse examine deux méthodes d'estimation parmi les plus importantes et les plus souvent utilisées - l'estimation du maximum de vraisemblance pour laquelle on supposée connue la distribution des variables (ou vectors) aléatoires et d'autre part l'estimation des moindres carrés où cette fois la distribution de variables aléatoires n'est pas nécessairement connue.

Quelques définitions et résultats préliminaires seront présentés dans le premier chapitre. Ceux-ci seront utiles dans les chapitres suivants. Le second chapitre portera sur les propriétés souhaitables des estimateurs qui seront rencontrés aux chapitres trois et quatre.

L'estimation du maximum de vraisemblance sera discutée au chapitre trois et l'emphase sera placée sur les propriétés assymptotiques des estimateurs. Le quatrième chapitre traitera de l'estimation des moindres carrés. Dans une section de ce chapitre, le problème d'estimabilité sera étudié. On trouvera de nombreuses références dans la bibliographie.

Department of Mathematics M.Sc.

ACKNOWLEDGEMENTS

My most sincere appreciation and thanks to Professor V. Seshadri for his undeniable guidance and support during the preparation of this thesis, and to the Canadian Commonwealth Scholarship Committee for financing my graduate studies.

I thank my fiancée Anthea whose frequent letters of encouragement have helped me greatly. I also thank my brothers Rupert and Horace and their families for providing me with an atmosphere conducive to studying.

Finally, I thank Mrs. Pamela Neblett for her patience, diligence and accuracy in the typing of this thesis.

TABLE OF CONTENTS

Page

1.15 Theorem: If A is a symmetric matrix, then there exists an orthogonal matrix P such that $PAP' = \Lambda$, where Λ is the diagonal matrix of characteristic roots of A and P the matrix of characteristic vectors.

1.16 The singular values of the matrix B $(m \times n)$ denoted by $Sg(B)$ are defined to be the positive square roots of the characteristic roots of the matrix $B'B$, i.e., $Sg(B) = +\left[CH(B'B)\right]^{\frac{1}{2}}$.

1.17 Let B $(m \times n)$ be an arbitrary real matrix. The rank of B, denoted by $r(B)$ is the maximal number of linearly independent rows (columns) of B.

1.18 If $n \leq m$ and $r(B) = n$, then B $(m \times n)$ is said to have full column rank, if $m \leq n$ and $r(B) = m$, B is said to have full row rank.

1.19 The matrix X $(n \times m)$ is said to be a right inverse of A $(m \times n)$ if $AX = I_m$. If Y $(n \times m)$ is such that $YA = I_n$, then Y is a left inverse of A.

1.20 Lemma: If A $(m \times n)$ is such that $r(A) = n$, then there exists a matrix X $(n \times m)$, called a right inverse of A, such that $AX = I_m$; one choice of X is $A'(AA')^{-1}$. If $r(A) = m$, then there exists a matrix Y such that $YA = I_n$; one choice of Y is $(A'A)^{-1}A'$.

1.21 A generalized inverse of A $(m \times n)$ is a matrix denoted by A^- of order $n \times m$ such that

$$A A^- A = A .$$

1.22 The Moore-Penrose generalized inverse of the matrix A $(m \times n)$ is a matrix G $(n \times m)$ such that the following four conditions are satisfied:

(1) $A G A = A$

(2) $G A G = G$

(3) $(AG)' = AG$

(4) $(GA)' = GA$.

The matrix G is called a $g_{1234}(A)$ and is denoted by A^+. If condition (1) is satisfied, G is called a weak generalized inverse of A and is denoted by $g_1(A)$. If conditions (1) and (2) are satisfied G is called a symmetric generalized inverse of A and is denoted by $g_{12}(A)$. If conditions (1) and (3) are satisfied G is called a least squares generalized inverse and is denoted by

$$g_{13}(A) \text{ or } A^-_{ls} .$$

1.23 Theorem: Let A $(m \times n)$ be an arbitrary real matrix; then there exists orthogonal matrices P $(m \times m)$ and Q $(n \times n)$ such that $P'AQ = \begin{bmatrix} D_r & 0 \\ 0 & 0 \end{bmatrix}$, where D_r is the diagonal matrix of singular values of A, P the matrix of

characteristic roots of $A'A$ and Q the matrix of characteristic roots of AA'.

1.24 Lemma: A^+ and $A_{\ell s}^-$ are unique.

1.25 Lemma: Using (1.23) we note that one representation of A^+ is

$$A^+ = Q\begin{bmatrix} D_r^{-1} & 0 \\ 0 & 0 \end{bmatrix} P' .$$

1.26 Theorem: For comformable matrices A and B and for any choice of the generalized inverses A^- and B^-, we have

$$r(A,B) = r(A) + r(\left[I-AA^-\right]B), \qquad (1.26.1)$$

$$= r(B) + r(\left[I-BB^-\right]A). \qquad (1.26.2)$$

$$r\begin{bmatrix} A \\ B \end{bmatrix} = r(A) + r(\left[B(I-A^-A)\right]), \qquad (1.26.3)$$

$$= r(B) + r(\left[A(I-B^-B)\right]). \qquad (1.26.4)$$

1.27 Lemma. $r(AB) \leq r(A)$, (1.27.1)

$r(AB) \leq r(B)$. (1.27.2)

1.28 Lemma: Let the matrix C $(m \times n)$, $n \leq m$, have full column rank n, and the matrix R $(n \times m)$, $n \leq m$, have full row rank n. Then for any matrix A $(n \times m)$, we have

$$r(A) = r(CA) = r(AR).$$

1.29 The trace of a matrix A $(n \times n)$ denoted by $tr(A)$ is defined to be the sum of the diagonal elements; alternatively the trace of A is the sum of the characteristic roots of A.

1.30 Lemma: If A $(n \times n)$ is an idempotent matrix of rank r, then $tr(A) = r$.

1.31 Lemma: Let $\alpha > 0$ and A $(m \times n)$ be a real matrix. Then

$$\lim_{\alpha \to 0} \left[A'A + \alpha I\right]^{-1} A' = (A'A)^{-} A'.$$

Proof: by (1.23), we have

$$A = P \begin{bmatrix} D_r & 0 \\ 0 & 0 \end{bmatrix} Q'$$

hence
$$A'A = Q \begin{bmatrix} D_r^2 & 0 \\ 0 & 0 \end{bmatrix} Q' \qquad (1.31.1)$$

and
$$A'A + \alpha I_n = Q \begin{bmatrix} D_r + \alpha I_r & 0 \\ 0 & \alpha I_{n-r} \end{bmatrix} Q' . \qquad (1.31.2)$$

We note that for all $\alpha > 0$, $A'A + \alpha I_n$ is invertible.

Thus
$$\left[A'A + \alpha I_n\right]^{-1} = Q \begin{bmatrix} \Lambda(\alpha) & 0 \\ 0 & \alpha^{-1} I_{n-r} \end{bmatrix} Q' \qquad (1.31.3)$$

where $\Lambda(\alpha) = \left[D_r + \alpha I_r\right]^{-1}$,

hence,
$$\left[A'A + \alpha I_n\right]^{-1} A' = Q \begin{bmatrix} \Lambda(\alpha) & 0 \\ 0 & \alpha^{-1} I_{n-r} \end{bmatrix} Q' Q \begin{bmatrix} D_r & 0 \\ 0 & 0 \end{bmatrix} P'$$

$$= Q \begin{bmatrix} \Lambda(\alpha) D_r & 0 \\ 0 & 0 \end{bmatrix} P', \qquad (1.31.4)$$

but $\lim_{\alpha \to 0} \Lambda(\alpha) = D_r^{-2}$,

hence

$$\lim_{\alpha \to 0} \left[A'A + \alpha I_n \right]^{-1} A' = Q \begin{bmatrix} D_r^{-1} & 0 \\ 0 & 0 \end{bmatrix} P'$$

$$= (A'A)^{-} A' .$$

Note: The above lemma can be generalized by replacing αI_n by $D(\alpha)$ a diagonal matrix of positive, not necessarily identical elements.

1.32 The density function $f(x)$ is a member of the exponential family if

$$f(x) = A(x)\, B(\theta) \exp\left[C(\theta)\, D(x) \right] , \quad a < x < b, \text{ where}$$

(1) θ is the vector of parameters,

(2) a or b does not depend on θ,

(3) $A(x)$ and $D(x)$ are continuous for $a < x < b$,

(4) $C(\theta)$ is a non-trivial continuous function of θ.

Examples of Exponential families are

(i) $\{ f(x): f(x) = \frac{1}{\theta} \exp(-\theta x), x \geq 0, \theta > 0 \}$

(ii) $\{ f(x): f(x) = \frac{1}{\Gamma(n)} \theta^n x^{n-1} \exp(-\theta x), x \geq 0, \theta > 0, n > 0 \}$.

1.33 Two subspaces are said to be virtually disjoint if the null vector is the only vector they have in common.

1.34 Let $A = \begin{bmatrix} E & F \\ G & H \end{bmatrix}$ with H nonsingular.

The Schur complement of E in A denoted by A/E is defined as;

$$A/E = E - F H^{-1} G .$$

CHAPTER II

DESIRABLE PROPERTIES OF ESTIMATORS

2.1 Introduction. The theory of mathematical estimation was founded by R. A. Fisher in a series of Papers (1921, 1925, 1934 and others). We shall discuss some of the important ideas developed by Fisher in this and a subsequent Chapter. We shall also classify and study the properties of various estimators and give some examples to strengthen the discussion.

2.2 The Problem. Suppose we have obtained a random sample from a population, whose distribution has a known mathematical form, but involves a certain number of unknown parameters. We would like to find estimates for the values of these unknown parameters in the parent population based on the information given by the sample.

In general, there will be an infinity of functions of the sample values that might be proposed as estimates of the parameters. We are faced with the following question; "how should we best combine the data to form these estimates?" This question leads to another, namely, "what do we mean by "best" estimates?"

We might answer impulsively that the best estimates will be the ones which are nearest to the true values of the parameters to be estimated. However, it must be borne in mind

that every suggested estimate is a function of the sample values, and hence must be regarded as an observed value of a certain random variable. We therefore see that we have no means of predicting the individual values assumed by the estimates in any given case, so that the goodness (in the sense of being best) of an estimate cannot be judged from individual values, but only from the distribution of the values which the estimate will assume in the long run, that is, from its sampling distribution.

When most of the "mass" in this distribution is concentrated in some small neighbourhood of the true value, there is a very small probability that the estimate will differ from the true value by more than a small amount. We express this probabilistically as $P\left[|t_n - \theta| > \varepsilon\right] < \delta$, where t_n is the estimator of the parameter θ , and δ is a small positive quantity depending on n and ε . From this point of view, an estimate will be "better" in the same measure as its sampling distribution shows a greater concentration about the true value.

We now list some desirable properties, some or all of which we would like our estimator to have:

(a) Consistency,

(b) Sufficiency,

(c) Unbiasedness,

(d) Minimum variance,

(e) Completeness,

(f) Efficiency.

We now define and give examples of statistics which satisfy the properties listed above.

2.3. Consistency

Definition: An estimator t_n computed from a random sample of n values is said to be a consistent estimator for the parameter θ, if for any given $\varepsilon > 0$ there exists an N and a δ such that

$$P(|t_n - \theta| \leq \varepsilon) \geq 1-\delta \qquad (1.3.1)$$

for all $n \geq N$; t_n is then said to converge in probability or to converge stochastically to θ.

Note: As defined, consistency is a large sample property.

Example

$$\text{Consider } dF(x) = \frac{1}{\sqrt{2\pi}} \exp -\frac{(x-\theta)^2}{2} \quad , -\infty < x < \infty$$

and let $x = (x_1, x_2, .. x_n)$ be a random sample of size n taken from this population. Then the sample mean $t = \sum_{i=1}^{n} \frac{x_i}{n}$ is consistent for the parameter θ, since, given $\varepsilon > 0$, and using the fact that $\sqrt{n}\,(t-\theta) \sim N(0,1)$, we have

$$P(\sqrt{n}|t - \theta| < \varepsilon\sqrt{n}) = \int_{-\varepsilon\sqrt{n}}^{\varepsilon\sqrt{n}} dF(x)$$

$$= \int_{-\varepsilon\sqrt{n}}^{\varepsilon\sqrt{n}} \exp(-\frac{t^2}{2}) \frac{dt}{\sqrt{2\pi}} \quad .$$

Clearly this probability can be made as close to one (1) as we please, by making n sufficiently large. For example, if $\varepsilon = 0.1$ and $n = 400$, we have,

$$P(20|t-\theta| \leq 2) = P(|t - \theta| \leq .1)$$

$$= 0.9544 .$$

It can be shown that the sample median, $x_{\frac{1}{2}}$, is also consistent for the parameter θ in the normal population and thus we note that consistent estimators are not necessarily unique. In cases where there are several consistent estimators it is usual to choose the estimator with the smallest variance.

In the case of a random sample of size n taken from a normal population with mean/median μ and variance σ^2, it can be shown that the sample mean/median are both asymptotically normal with common expected value μ and variances $\frac{\sigma^2}{n}$ and $\frac{\pi\sigma^2}{2n}$ respectively. Thus we see that the sample mean will be preferred to the sample median as an estimator for the population mean, since the sample mean has asymptotically smaller variance than the sample median.

2.4. Suffciency

Many decision theoretic problems can be significantly simplified by a suitable reduction in the complexity of the data. A suitable form of reduction called the criterion of sufficiency was introduced by R. A. Fisher (1925). There he defined a sufficient statistic as one which effectively summarizes all the relevant information supplied by the data.

The following definition offers a small extension

to Fisher's concept of sufficiency.

Definition: A statistic is said to satisfy the criterion of sufficiency when no other statistic which can be calculated from the data provides any additional information as to the value of the parameter to be estimated.

A more rigorous definition is as follows:

Definition: Let X denote a random variable (vector) whose distribution depends on a parameter $\theta \in \Theta$, Θ being the parameter space. A real (vector) function T of X is said to be sufficient for the parameter θ if the conditional distribution of X given $T = t$ is independent of θ almost everywhere (t).

The following theorem, due to Fisher, gives us a necessary and sufficient condition for a statistic to be sufficient.

Fisher's Factorization Theorem:

Let X be a random variable whose p.d.f. (p.m.f) $f(x,\theta)$ depends on the parameter $\theta \in \Theta$. A function $T = t(x)$ is sufficient for θ if and only if the frequency function factors into a product of the two functions g and h, where g is a function of $t(x)$ and θ , and h is a function of x only, i.e.,

$$f(x,\theta) = g(t(x),\theta)\, h(x) .$$

Proof (if) Suppose that $T = t(x)$ is sufficient for θ then,

$$f(x,\theta) = P_\theta(X = x)$$
$$= P_\theta(X = x, T = t(x))$$
$$= P_\theta(T = t(x))\ P(X = x \mid T = t(x)) \qquad (2.4.1)$$

provided that $P(X = x \mid T = t(x))$ exists and is well defined. Hence for the x for which $f(x,\theta) = 0$ for all θ

we define

$$h(x) = 0$$

and for the x for which $f(x,\theta) > 0$ for some θ

we put

$$h(x) = P(X = x \mid T = t(x)) \qquad (2.4.2)$$

which is independent of θ, so that putting

$$g(t(x),\theta) = P_\theta(T = t(x)) \qquad (2.4.3)$$

gives us the required factorization.

Proof (only if) Suppose

$$f(x,\theta) = P_\theta(X = x)$$
$$= P_\theta(T = t(x))\ P(X = x \mid T = t(x)) \qquad (2.4.4)$$

holds. Fix t_0 for which $P_\theta(T = t_0) > 0$ for some $\theta \in \Theta$. Then

$$P_\theta(X = x \mid T = t_0) = \frac{P_\theta(X = x, T = t_0)}{P_\theta(T = t_0)} \qquad (2.4.5)$$

The numerator of (2.4.5) is zero for all θ whenever $t(x) \neq t_0$ and is equal to $P_\theta(X = x)$, whenever $t(x) = t_0$.

We can write the denominator of (2.4.5) as

$$P_\theta(T = t_0) = \sum_{[x:t(x) = t_0]} P_\theta(X = x)$$

$$= \sum_{[x:t(x) = t_0]} g\big(t(x),\theta\big).h(x) \quad . \qquad (2.4.6)$$

Hence

$$P_\theta(X = x \mid T = t_o) = 0 \qquad t(x) \neq t_0$$

$$= \frac{g(t_0,\theta).h(x)}{g(t_0,\theta).h(x_1)}, \qquad t(x) = t_0$$

where $h(x_1) = \sum_{[x':t(x') = t_0]} h(x')$.

Cancelling $g(t_0,\theta)$, we have that,

$P_\theta(X = x \mid T = t_0)$ is independent of θ, for all t_0 and θ for which it is defined.

We shall see later that using the criterion of sufficiency in conjunction with that of completeness, we can construct uniformly minimum variance unbiased estimators (UMVUE) for estimable functions. We give here two examples of how to find a sufficient statistic using the Fisher Factorization Criterion.

Example I. Let $x = (x_1, x_2, \ldots, x_n)$ be a random sample from a Poisson Population with parameter $\theta > 0$. Then the likelihood function is

$$L(\theta|x) = \prod_{i=1}^{n} \frac{\theta^{x_i}}{x_i!} \exp(-\theta)$$

$$= \frac{1}{x_1! x_2! \ldots x_n!} \cdot \theta^{\sum_{i=1}^{n} x_i} \exp(-n\theta)$$

$$= h(x)\, g(t(x), \theta)$$

Where $h(x) = \left[\prod_{i=1}^{n} x_i!\right]^{-1}$ and $t(x) = \sum_{i=1}^{n} x_i$,

thus $t(x)$ is sufficient for θ. It is easily seen that $\frac{t(x)}{n}$ is a consistent and unbiased estimator for θ.

Example II. Let $x = (x_1, x_2, \ldots, x_n)$ be a random sample from a distribution with pdf

$$f(x,\theta) = G(\theta)M(x) \quad \text{if } 0 < x < \infty$$

$$= 0 \quad \text{elsewhere,}$$

where $\left[G(\theta)\right]^{-1} = \int_0^{\theta} M(x)dx$.

The likelihood function is:

$$L(\theta|x) = \prod_{i=1}^{n} G(\theta)M(x_i) \quad \text{if all } x_i \in (0,\theta)$$

We can then write $L(\theta|x)$ as

$$L(\theta|x) = q(0, y_1) \prod_{i=1}^{n} M(x_i)\, q(y_n, \theta) \left[G(\theta)\right]^n$$

$\forall\ x_1, x_2, \ldots, x_n$, where y_1 is the first order statistic,

y_n is the n^{th} order statistic, and

$$q(a,b) = 1 \quad \text{if } a < b$$

$$= 0 \quad \text{if } a > b \ .$$

If we now let $h(x) = q(0,y_1) \prod_{i=1}^{n} M(x_i)$

and $g(t(x),\theta) = q(y_n,\theta) . \left[G(\theta)\right]^n$,

we see that the n^{th} order statistic y_n is sufficient for the parameter θ.

Note: When $G(\theta) = \theta^{-1}$ and $M(x) = 1$ we have the uniform distribution. Another specific example of this form is the distribution with $M(x) = nx^{n-1}$ and $G(\theta) = \theta^{-n}$.

2.5. Unbiasedness

Definition: An estimator t_n of a parameter θ is said to be unbiased if the expected value of t_n is equal to θ, i.e. if $E[t_n] = \theta$; otherwise the estimator is said to be biased with a certain bias, $b(\theta)$.

We mentioned earlier that it is very desirable to choose a consistent statistic as an estimator of the population parameter. In general, such estimators are biased for the parameter. However, in some cases it is easy to change these biased estimators into unbiased estimators by the introduction of a constant scale and or shift factor as the case may be. In other instances when the expected value is a very complicated function of the parameter it is not so easy to determine these scale and shift factors. However, Quenouille (1956) has suggested a rather elegant method for overcoming this difficulty in many situations. The interested reader is referred to "The Theory of Advanced Statistics, Vol. II" by M. G. Kendall and A. Stuart.

We now give an example in which a single scale factor is necessary to change a biased estimator into an unbiased estimator.

Example: Let $x_1, x_2, \ldots, x_n$ be a sample of size n taken from a univariate normal population with known mean μ and unknown variance σ^2. Then

$$S = \sum_{i=1} (x_i-\mu)^2/n$$

is the M.L.E for σ^2. It is however biased, since $E[S] = \frac{n-1}{n}\sigma^2$. We therefore see that $\frac{nS}{n-1}$ is unbiased for σ^2.

2.6 Minimum Variance

We mentioned in the discussion of consistent estimators that when there are two or more consistent estimators for a parameter the desirable estimator is the one with the smallest variance. This leads us to a very important class of estimators, namely, minimum variance estimators. The concept of minimum variance has been used since the days of Gauss and Laplace. However, in relatively recent times, Cramér, Rao and Battacharrya have shown that lower bounds for the variance exist. We shall deal with minimum variance estimators which are also consistent and unbiased.

Definition: A consistent estimator t_n of a parameter θ is a minimum variance unbiased estimator if

$$\text{(i)} \quad E[t_n] = \theta$$

and

$$\text{(ii)} \quad \lim_{n\to\infty} \text{Var}(t_n) \leq \lim_{n\to\infty} \text{Var}(t_n')$$

for all consistent unbiased estimators t_n'.

We now state and prove a theorem on minimum variance bounds and discuss methods of determining such bounds.

Theorem: If $L(\theta)$ and t satisfy the regularity assumptions RA1-RA5 and if $E_\theta[t] = \tau(\theta)$

then $\text{Var}(t) \geq \frac{\partial}{\partial\theta}\tau(\theta)\, J(\theta)^{-1} \frac{\partial}{\partial\theta'}\tau(\theta)$. (2.6.1)

The univariate analogue of this theorem is known as the Cramér-Rao theorem and the quantity on the right of inequality (2.6.1) is called the Cramér-Rao (sometimes the Frechet-Cramér-Rao) lower bound.

Proof of the Theorem: Let X_1 and X_2 be arbitrary random vectors with means μ_1 and μ_2 covariance matrices Σ_{11} and Σ_{22} respectively and cross covariance matrix Σ_{12}, we assume that the matrix Σ_{22} is positive definite.

Using 1.32, we have the following identity

$$\Sigma_{11} = \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21} + \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21} \qquad (2.6.2)$$

$\Sigma_{11} - \Sigma_{12}\Sigma_{22}\Sigma_{21}$, is the schur complement of Σ_{11} in $\Sigma = \begin{bmatrix}\Sigma_{11} & \Sigma_{12}\\ \Sigma_{21} & \Sigma_{22}\end{bmatrix}$

Setting $X_1 = t$ and $X_2 = S$, the score vector, we have $\mu_1 = \tau(\theta)$ and $\mu_2 = 0$,

$$\Sigma_{12} = \Sigma_{t,s} = \frac{\partial}{\partial\theta}\tau(\theta) = \Sigma_{21}$$

$$\Sigma_{22} = \Sigma_{s,s} = J(\theta),\ \Sigma_{11} = \Sigma_{t,t}\ .$$

Thus

$$\Sigma_{11} = \Sigma_{tt} = \frac{\partial}{\partial\theta}\tau(\theta)\, J(\theta)^{-1}\frac{\partial}{\partial\theta'}\tau(\theta) + \text{Var}\left[t - \tau(\theta) - \Sigma_{12}\Sigma_{22}^{-1}S\right]$$

Hence, $\Sigma_{t,t} \geq \frac{\partial}{\partial\theta}\tau(\theta)\, J(\theta)^{-1} \frac{\partial}{\partial\theta'}\tau(\theta)$,

since a covariance matrix is non-negative definite. We have equality if and only if

$$t = \tau(\theta) + \Sigma_{12}\Sigma_{22}^{-1} S$$

almost everywhere (t). Q.E.D.

Note: For non-negative definite matrices A and B, $A \geq B$ means that A-B is non-negative definite.

Example: Let $X \sim N_p(\mu,\Sigma)$.

Suppose a random sample of size n is taken from this normal population and suppose also that Σ is positive definite. Then,

$$\bar{X} = \sum_{i=1}^{n} x_i/n$$

is sufficient for μ, and has minimum variance.

Proof:

$$L(\mu,\Sigma|X) = (2\pi)^{\frac{-np}{2}}|\Sigma|^{-n/2}\exp\left[\sum_{i=1}^{n}(x_i-\mu)'\Sigma^{-1}(x_i-\mu)\right].$$

Therefore,

$$\text{Log } L(\mu,\Sigma|X) = -\frac{nP}{2}\log 2\pi - \frac{n}{2}\log|\Sigma| + \sum_{i=1}^{n}(x_i-\mu)'\Sigma^{-1}(x_i-\mu),$$

$$\frac{\partial \ell}{\partial\mu} = \frac{\partial}{\partial\mu}\sum_{i=1}^{n}(x_i-\mu)'\Sigma^{-1}(x_i-\mu)$$

$$= \sum_{i=1}^{n}\frac{\partial}{\partial\mu}(x_i-\mu)'\Sigma^{-1}(x_i-\mu)$$

$$= \sum_{i=1}^{n}\left[-2\Sigma^{-1}x_i + 2\Sigma^{-1}\mu\right]$$

$$= -2\Sigma^{-1}\left[\sum_{i=1}^{n}(x_i-\mu)\right]$$

$$= -2n\Sigma^{-1}(\bar{x}-\mu) \quad .$$

Hence,

$$\frac{\partial \ell}{\partial \mu} = 0 \quad \Rightarrow \quad \bar{x} = \mu .$$

Thus $\bar{x}$ is the maximum likelihood estimator for μ and the Cramér-Rao bound Σ/n is attained by this estimator, a fact easy to verify.

Note that maximum likelihood estimators attain the Cramér-Rao lower bound. Furthermore in order to compute the Cramér-Rao lower bound, it is not necessary to find the variance of the maximum likelihood estimator explicitly. We just set $\frac{\partial \ell}{\partial \mu} = 0$, divide by the constant, and the coefficient of $(t - \mu)$ will be the Cramér-Rao bound; here, t is the maximum likelihood estimator of μ.

2.7 Completeness

The concept of completeness of a parametric family was introduced by Lehmann and Scheffé (1950). Completeness confers the uniqueness property on the class of sufficient estimators. This property taken in conjunction with that of sufficiency will be used to determine uniformly minimum variance unbiased estimators for estimable functions of a parameter θ.

Definition: A parametric family of distributions $f(x,\theta)$, depending on the value of the parameter θ, is said to be complete if, for $h(x)$ any statistic independent of the parameter θ, the expected value of $h(x)$ being equal to zero implies that $h(x) = 0$ almost everywhere (x).

If the expected value of $h(x) = 0$ implies that $h(x) = 0$ only for bounded $h(x)$, then $f(x, \theta)$ is said to be boundedly complete.

Definition: A statistic $t = \phi(x)$ is complete, if the family of distribution functions of t is complete.

We list here with proofs a few standard examples of complete families of distribution functions.

2.7.1 The family of binomial distributions with

p.m.f $\binom{n}{x} p^x (1-p)^{n-x}, \quad x = 0,1,\ldots,n$

$0 < p < 1$, is complete, since, if

$$E_p[h(x)] = \sum_{x=0}^{n} h(x)\binom{n}{x}p^x (1-p)^{n-x} = 0, \forall\, p,\ 0 \leq p \leq 1.$$

then this n^{th} degree polynomial in p must be the unique zero polynomial. That is, the coefficients of each power of p must be zero.

Now,

$$\sum_{x=0}^{n} h(x)\binom{n}{x} p^x (1-p)^{n-x}$$

$$= \sum_{x=0}^{n} \sum_{j=0}^{n-x} (-1)^j \binom{n}{x}\binom{n-x}{j} h(x)\ p^{x+j}$$

$$= 0 \ .$$

We therefore have

$$1 = p^0 \Rightarrow x = 0,\ j = 0$$

hence, $h(0) = 0$.

$p^1 \rightarrow x = 0,\ j = 1 \quad \text{or} \quad x = 1,\ j = 0$

hence $\binom{n}{1} h(0) + \binom{n}{0} h(1) = 0$

or $\quad h(1) = 0, \quad \text{since } h(0) = 0 \cdot$

Continuing in this manner we see that

$$h(0) = h(1) = \ldots = h(n) = 0 \cdot$$

Q.E.D.

2.7.2. The family of all Poisson distributions with

p.m.f. $\frac{\theta^x}{x!}\exp(-\theta), \quad x = 0,1 \ldots \quad \theta > 0$ is complete.

Proof:

$$f(x,\theta) = \frac{\theta^x}{x!}.\exp(-\theta), \quad x = 0,1,2,\ldots \quad \theta > 0.$$

If $h(x)$ is any function of the data, we have

$$E_\theta\left[h(x)\right] = \sum_{x=0}^{\infty} h(x)\frac{\theta^x}{x!}\exp(-\theta) = 0 \cdot$$

If we expand $\exp(-\theta)$ as a Taylor series about zero, we get

$$E_\theta\left[h(x)\right] = \sum_{x=0}^{\infty} h(x).f(\theta) \equiv 0$$

hence, $h(x) = 0$ almost everywhere (x), because of the uniqueness of the Taylor series expansion of $f(\theta)$.

The family of normal distributions with mean zero and variance $\sigma^2 > 0$ is not complete, since any odd function $h(x)$ has expectation zero. We now state and prove two theorems, the first of which gives us a means of improving on an arbitrary unbiased estimator while the second theorem gives us two ways in which we can determine a uniformly minimum variance estimator.

2.7.3. Completeness and the Rao-Blackwell Theorem

Theorem: If the statistic t is sufficient for the parameter θ and $E_\theta[t'] = g(\theta)$, then,

$$t'' = E_\theta[t' \mid t]$$

is such that t'' is an unbiased estimator of $g(\theta)$; t'' is therefore θ-free and is a function of t only, say, $t'' = h(t)$ such that

$$\text{Var}(t'') \leq \text{Var}(t')$$

with equality holding if and only if t' and t are dependent with probability one.

Proof: t sufficient for θ implies that the distribution of $t' \mid t$ is θ-free, hence, $E_\theta[t' \mid t]$ is θ-free; so that, $E_\theta[t' \mid t]$ is a statistic.

Now, $t' = E_\theta[t' \mid t] + \varepsilon_{t';t}$, where $\varepsilon_{t',t}$ is the error due to the regression of t' on t and $E_\theta[t' \mid t] = h(t) = t''$ is the regression function of t' on t. Also,

$$E_\theta[h(t)] = E_\theta[t''] = g(\theta), \text{ say.}$$

Furthermore, since $h(t)$ and $\varepsilon_{t';t}$ are uncorrelated (this is immediate from regression theory), we have

$$\text{Var}(t') = \text{Var}(t'') + \text{Var}(t' - h(t)),$$

hence,

$$\text{Var}(t') \leq \text{Var}(t'')$$

with equality holding if and only if $t' = h(t)$. (i.e. t' and

t are functionally dependent) with probability one.

We note that the Rao-Blackwell theorem gives us a way of improving an arbitrary unbiased estimator t' of $g(\theta)$ when a sufficient statistic t for θ exists and is functionally independent of θ; t'' calculated from t' and t by the above method is unbiased for the parameter θ and it has variance less than the variance of t', for all sufficient statistics t. We also note that in seeking a uniformly minimum variance unbiased estimator of $g(\theta)$, we only need consider unbiased estimators which are functions of the sufficient statistic t, since any other estimator for θ cannot be a uniformly minimum variance unbiased estimator. The above idea will now be used in the second theorem.

2.7.4 The Lehmann-Scheffé Theorem.

We have previously mentioned that completeness confers the uniqueness property on sufficient statistics. The following theorem, due to Lehmann and Scheffé, characterizes this uniqueness property.

Theorem. If the statistic t is complete and sufficient for a parameter θ, and $g(\theta)$ is an estimable function of θ, then the function $g(\theta)$ has a uniformly minimum variance unbiased estimator h(t), a function of t only, which can be determined essentially in either of the following two ways.

(i) If t' is any unbiased estimator of the function $g(\theta)$, which is functionally independent of the statistic t, then

$$h(t) \equiv E_{\theta}[t' | t]$$

(ii) $h(t)$ is the solution of the equation

$$E_\theta[h(t)] = g(\theta) .$$

Proof. Define $U = U_t \bigcup \bar{U}_t$, where U_t is the set of all unbiased estimators of $g(\theta)$ which are functionally dependent on t and $\bar{U}_t$ is the set of all unbiased estimators of $g(\theta)$ which are functionally independent of t. U is the set of unbiased estimators. Then, $U_t \bigcap \bar{U}_t = \phi$, ϕ denotes the null set. Since $g(.)$ is an estimable function and $U \neq \phi$, U_t and $\bar{U}_t$ cannot both be empty.

(i) Assume $\bar{U}_t$ is not empty and let $t' \in \bar{U}_t$, then, by the Rao-Blackwell theorem,

$$t'' = E_\theta[t'|t] = h(t), \quad \text{say},$$

is an unbiased estimator for $g(\theta)$, and $h(t)$ is an element of U_t which has uniformly smaller variance than t'. Moreover, since t is complete for θ, $h(t)$ is an essentially unique member of U_t and is determined by the relationship

$$E_\theta[h(t)] = g(\theta) .$$

Thus $h(t)$ is a uniformly minimum variance unbiased estimator of $g(\theta)$.

(ii) Suppose U_t is not empty, then as in (i) U_t contains essentially only one element $t'' = h(t)$, which is determined by

$$E_\theta[h(t)] = g(\theta)$$

and hence t'' is a uniformly minimum variance unbiased estimator for $g(\theta)$. We are now in a position to prove the

uniqueness of a uniformly minimum variance unbiased estimator for a parametric function.

Proof: Suppose there are two such estimators of $g(\theta)$, t_1 and t_2, say. Then $t_3 = (t_1 + t_2)/2$ is also unbiased for $g(\theta)$.

Let V by the common variance of t_1 and t_2

Then,

$$\text{Var}(t_3) = (2V + 2\,\text{Cov}(t_1,t_2))/4$$

$$= (V + \text{Cov}(t_1,t_2))/2 \ .$$

Now, by the Cauchy-Schwarz inequality

$$\text{Cov}(t_1, t_2) \leq \sqrt{(\text{Var}\ t_1\ \ \text{Var}\ t_2)} = V \ .$$

Thus $\text{Var}(t_3) \leq V$, which contradicts the assumption that t_1 and t_2 are minimum variance unbiased estimators.

Hence, $\text{Cov}(t_1, t_2) = V$

and $\text{Var}(t_3) = V$.

This can only be so if

$$t_1 - g(\theta) = k(\theta)\,(t_2 - g(\theta)) \ ,$$

i.e., if t_1 is proportional to t_2, but then,

$$\text{Cov}(t_1, t_2) = k(\theta).V = V$$

hence $k(\theta) = 1$, and so $t_1 \equiv t_2$.

Q.E.D.

We now give an example of how to determine the uniformly minimum variance estimator.

Example: Let $X_1, X_2, \ldots, X_n$ be a random sample from a uniform distribution on $(0, \theta)$, $0 < \theta < \infty$, let $t = \max (X_1, X_2, \ldots X_n)$. It can be shown that t is complete and sufficient for the parameter θ.

We are to determine a uniformly minimum variance unbiased estimator of θ^r, $r \neq -n$.

By the Lehmann-Scheffé theorem some function of the statistic t will be such an estimator.

Assume t^r is an estimator of θ^r.

then,

$$E_\theta\left[t^r\right] = \int_0^\theta t^r g(t)dt$$

$$= \int_0^\theta t^r . \frac{nt^{n-1}}{\theta^n} \, dt$$

$$= \frac{1}{\theta^n}\left[\frac{n}{n-r} t^{n-r}\right]_0^\theta$$

$$= \frac{n}{n-r} \theta^r \, .$$

Thus $\frac{n-r}{n} t^r$ is unbiased for θ^r and will be the uniformly minimum variance unbiased estimator.

2.8 Efficiency

We noticed earlier that the sample mean $\bar{x}$ and median $x_{\frac{1}{2}}$ were two consistent estimators for the population mean of a normal distribution with mean μ and variance σ^2, and that these two estimators were both asymptotically normal with the same mean μ and variance σ^2/n and $\pi\sigma^2/2n$ respectively.

There is a large and important class of consistent estimators whose sampling distributions are asymptotically normal and for which the variance decreases as the sample size increases, so that the properties of such estimators can be characterized by the variance and bias. The bias will not be considered, since it must tend to zero as the sample size increases, or in the first case we could have made a correction for the bias, as noted earlier. We are not so fortunate when it comes to the variance, since the variance is invariant under a constant shift. We therefore seek a criterion which will enable us to choose an estimator from among competing ones.

The concept of efficiency as defined by R. A. Fisher (1921) requires that the fixed value to which n times the estimator tends, shall be as small as possible. An efficient estimator will be such that this property is satisfied.

If the variance of any efficient estimator is known, then, the efficiency of any other estimator of the same

parameter is calculated as the ratio of the variance of the efficient estimator to that of the other estimator.

Definition: In the class of all unbiased estimators of a parameter θ, $\hat{\theta}$ is efficient if

$$\text{Var}(\hat{\theta}) \leq \text{Var}(\hat{\theta}')$$

that is, $\hat{\theta}$ is efficient if it is a uniformly minimum variance unbiased estimator.

Example. Consider again a normal population with mean μ and variance σ^2. Let $x_1, x_2, \ldots x_n$ be a random sample of size n chosen from this population. As mentioned earlier, the sample mean $\bar{x}$ and the sample median $x_{\frac{1}{2}}$ are both unbiased consistent estimators for the population mean μ. However, the variance of the sample mean is σ^2/n and that of the sample median is $\pi\sigma^2/2n$. Hence the efficiency of the median is

$$\frac{\sigma^2/n}{\pi\sigma^2/2n} = \frac{2}{\pi} \simeq .637 .$$

This means that from 1000 observations the sample median will produce a value which will be as accurate as that produced by the sample mean for only 637 observations, and therefore as mentioned in the discussion of consistency, the sample mean would be a better choice as an estimator for the population mean in the normal population.

Sometimes, it is more difficult to calculate an efficient estimator than an inefficient one, and the labour involved in calculating this efficient estimator might out-

weigh the cost of taking additional observations needed for the inefficient estimator. In such cases, the additional observations may be taken and the inefficient estimator chosen as the estimator for the population parameter.

2.8.1. Properties of efficient estimators.

(i) The correlation between any two statistics, both efficient estimators of the same parameter θ, tends to $+1$, as the sample size increases. For if t_1 and t_2 are two such statistics whose variances are σ^2/n, say, and the correlation is r, we choose $t_3 = (t_1 + t_2)/2$. Then, the statistic t_3 will be an efficient estimator of the parameter θ. The variance of t_3 is $(1 + r)\sigma^2/n$, and this, by hypothesis, must be at least σ^2/n. Hence, r must be at least $+1$, but r is at most $+1$, therefore $r = +1$. This property implies that for large samples all efficient statistics (estimators) are equivalent.

(ii) If we relax the constraint of unbiasedness in the definition of efficiency and replace the variance by the mean square error, we note that since maximum likelihood estimators have asymptotically smaller variances (see Kendall and Stuart, Vol.II) they are therefore efficient.

CHAPTER III

MAXIMUM LIKELIHOOD ESTIMATION

3.1 Introduction. One of the most important methods of estimation of population parameters is the method of Maximum Likelihood. This method of estimation was formally introduced by Fisher, R.A. in a short paper in 1912, and was further developed in a series of papers (Fisher (1922, 1925, 1934)) by the same author. Several other authors have since made important contributions to the theory.

Consider a random sample $X = (x_1, x_2, \ldots, x_n)'$ taken from a population with parameter θ, and known probability density (mass) function (p.d.f.) f. The joint probability of the observations regarded as a function of the unknown parameter θ is called the likelihood function, and is denoted by $L(X \mid \theta)$, or simply by $L(\theta)$, and is given by

$$L(\theta) = \prod_{i=1}^{n} f(x_i \mid \theta), \quad i=1,2,\ldots,n.$$

We note here that since $L(\theta)$ is the product of p.d.f.'s (p.m.f.'s) it is necessarily positive.

The method of Maximum Likelihood in its simplest form leads us to take as estimate of the parameter θ, that value $\hat{\theta}$ which lies within the range of admissible values of the parameter $\hat{\theta}$ and makes the likelihood function $L(\theta)$ as large as possible,

that is,

$$L(\hat{\theta}) \geq L(\theta) \quad \text{for all values of } \theta \in \Theta$$

the parameter space, and $\hat{\theta} \in \Theta$.

If we assume that $L(\theta)$ is a twice differentiable function (in the range of θ) with respect to θ, then under suitable conditions we can use calculus to find $\hat{\theta}$. Namely, we solve the following constrained system

$$\frac{\partial}{\partial \theta} L(\theta) = 0, \tag{3.1.1}$$

subject to

$$\frac{\partial^2}{\partial \theta \partial \theta'} L(\hat{\theta}) < 0^{*}. \tag{3.1.2}$$

We note that the condition (3.1.2) is only a sufficient condition for maximality of $\hat{\theta}$, and that it is not a necessary condition. For consider

$$g(\theta) = -\theta_1^4 - \theta_2^4 \quad , \qquad \theta = (\theta_1, \theta_2)' \ ,$$

then

$$\frac{\partial}{\partial \theta} g(\theta) = (-4\theta_1^3, -4\theta_2^3) \ ,$$

and

$$\frac{\partial^2}{\partial \theta \partial \theta'} g(\theta) = -12 \begin{bmatrix} \theta_1^2 & 0 \\ 0 & \theta_2^2 \end{bmatrix} .$$

$\theta = (0, 0)'$ can be shown to be the point of maximum of $g(\theta)$, however

$$\frac{\partial^2}{\partial \theta \partial \theta'} g(\theta) \Big|_{\theta = (0,0)'}$$

* The matrix inequality $<$ indicates negative definiteness.

is not negative definite.

In many situations the system of equations obtained by setting $\frac{\partial L(\theta)}{\partial \theta} = 0$ is quite complicated and rather difficult to solve. In some of these cases it is much easier to work with the logarithm of the likelihood function denoted by $\ell(\theta)$. This is possible because a positive function and its logarithm attain their maxima (minima) at the same point. Thus we solve the following constrained system;

$$\frac{\partial}{\partial \theta} \ell(\theta) = 0, \tag{3.1.3}$$

subject to

$$\frac{\partial^2}{\partial \theta \partial \theta'} \ell(\hat{\theta}) < 0 . \tag{3.1.4}$$

We know from (1.7) that

$$\frac{\partial}{\partial \theta} \ell(\theta) = S,$$

and from (1.8) that

$$- E\left[\frac{\partial^2}{\partial \theta \partial \theta'} \ell(\theta)\right] = J(\theta).$$

3.2 Regularity Assumptions

Before discussing iterative methods for the solution of the likelihood equation, we mention here some regularity assumptions which we place on the likelihood function $L(\theta)$, and discuss the properties of the score vector, $S(\theta)$, and the information matrix, $J(\theta)$.

RA1. $\theta \in \Theta_\ell$, where Θ_ℓ is a subset of $\mathbb{R}^\ell$ (Euclidean ℓ-space) possibly equal to $\mathbb{R}^\ell$.

RA2. For all samples X, $\frac{\partial L(\theta)}{\partial \theta_j}$, $j = 1,\ldots,\ell$, exists for all $\theta \in \Theta_\ell$ except perhaps for a set of points in Θ_ℓ with probability measure zero.

RA3. $\int_X L(\theta)\, dx$ (or $\Sigma L(\theta)$) can be differentiated under the integral (summation) sign with respect to θ_j, $j = 1, \ldots, \ell$.

RA4. $\int_X t_i\, L(\theta)\, dx$ (or $\Sigma\, t_i\, L(\theta)$) can be differentiated under integral (summation) sign with respect to θ_j, $j = 1, \ldots, \ell$, where (t_i), $i=1, \ldots, m$ is a sufficient statistic for the parameter $\tau(\theta)$.

RA5. The score vector $S(\theta) = \frac{\partial}{\partial \theta}\, \ell(\theta)$ has a positive definite covariance matrix for each $\theta \in \Theta_\ell$.

3.3. Properties of the Score Vector

(1) Let x_i $(m_i \times 1)$ $i = 1, \ldots, q$ be independent vectors, where the distribution of each x_i involves the same θ. Let $X' = (x_1', x_2', \ldots, x_q')$. Denote the score vectors of x_i and X by $S_i(\theta)$ and $S(\theta)$ respectively. Then

$$S(\theta) = \sum_{i=1}^{q} S_i(\theta) \quad . \tag{3.3.1}$$

Proof: $L(\theta,X) = \prod_{i=1}^{q} L(\theta,x_i)$.

Hence, $\log L(\theta,X) = \sum_{i=1}^{q} \log L(\theta,x_i)$.

Whence, $\frac{\partial}{\partial\theta_j} \log L(\theta,X) = \sum_{i=1}^{q} \frac{\partial}{\partial\theta_j} \log L(\theta,x_i)$, $j = 1,\ldots,\ell$.

Thus

$$S(\theta) = \sum_{i=1}^{q} S_i(\theta).$$

(ii) For all values of $\theta \in \Theta_\ell$, the expected value of the score vector is zero.

Proof: Since $L(\theta)$ is a probability density, we have

$$\int_X L(\theta)dx = 1 .$$

Hence, using RA2 and RA3, we see that

$$\int_X \frac{\partial}{\partial\theta_j} L(\theta)\, dx = 0, \quad j = 1,\ldots,\ell .$$

But,

$$\int_X \frac{\partial}{\partial\theta_j} L(\theta)\, dx = \int_X \frac{1}{L(\theta)} . \frac{\partial L(\theta)}{\partial\theta_j} . L(\theta)dx$$

$$= \int_X S_j(\theta).L(\theta)dx .$$

Hence $E\left[S(\theta)\right] = 0$.

(iii) The covariance matrix $\Sigma_{t,S}$ between S and any unbiased estimator t is given by

$$\Sigma_{t,S} = \frac{\partial}{\partial\theta}\tau(\theta),$$

where $\tau(\theta) = E[\,t\,]$. In particular, for $\tau(\theta) = \theta$, we have $\Sigma_{t,S} = I$.

Proof. $$E[\,t\,] = \int_X t.L(\theta)dx = \tau(\theta).$$

Thus using RA3 and RA4, we have

$$\int_X t \frac{\partial}{\partial\theta} L(\theta)\, dx = \frac{\partial}{\partial\theta}\tau(\theta),$$

i.e., $$\int_X t \frac{1}{L(\theta)} \frac{\partial}{\partial\theta} L(\theta) \; L(\theta)\, dx = \frac{\partial}{\partial\theta} \tau(\theta) ,$$

or $$\int_X tS(\theta)L(\theta)\, dx = \frac{\partial}{\partial\theta} \tau(\theta)$$

$$= \Sigma_{t,S} .$$

3.4 Properties of the Information Matrix

(i) Additivity:

If $J_1(\theta)$, $J_2(\theta)$, ..., $J_n(\theta)$ are the information matrices for n independent random experiments, then the information matrix $J(\theta)$ for the combined experiment is given by

$$J(\theta) = \sum_{i=1}^{n} J_i(\theta) .$$

If $X' = (Z_1', Z_2', \ldots, Z_n')$, where X is the ordered observations from the combined experiment, and Z_1, Z_2, ..., Z_n is a random sample from a distribution with p.d.f. (p.m.f.) $f(x,\theta)$, then

$$J(\theta) = nJ_0(\theta),$$

where $J_0(\theta) = E\left[\frac{\partial}{\partial\theta} \log f \left[\frac{\partial}{\partial\theta} \log f\right]'\right]$.

$J_o(\theta)$ is called the information matrix per unit observation.

Proof

Let $S_i(\theta)$ $i = 1,\ldots,n$ be the score vector for the ith experiment, and let $S(\theta)$ be the score vector for the combined experiment.

Then $J(\theta) = E\left[S(\theta)\, S(\theta)'\right]$

$$= E\left[\sum_{i=1}^{n} S_i(\theta) . \sum_{j=1}^{n} S_j(\theta)'\right]$$

$$= E\left[\sum_{i=1}^{n} S_i(\theta)S_i(\theta)' + 2\sum_{i<j}^{n} S_i(\theta)S_j(\theta)'\right]$$

$$= \sum_{i=1}^{n} E\left[S_i(\theta)S_i(\theta)'\right]$$

$$= \sum_{i=1}^{n} J_i(\theta), \text{ since } S_i(\theta) \text{ and } S_j(\theta) \text{ are}$$

independent for $i \neq j$, being functions of Z_i and Z_j respectively, with Z_i and Z_j independent by hypothesis.

(ii) If in addition to the assumptions RA1 to RA5 we assume that $\frac{\partial^2}{\partial\theta\partial\theta'} L(\theta)$ exists almost everywhere (x) for all $\theta \in \Theta_\ell$ and if $\int_X L(\theta)dx$ can be differentiated twice under the integral sign, then

$$J(\theta) = E\left[-\frac{\partial^2}{\partial\theta\partial\theta'} L(\theta)\right] .$$

Proof:

$$\frac{\partial}{\partial\theta} \log L(\theta) = \frac{1}{L(\theta)} \frac{\partial}{\partial\theta} L(\theta) ,$$

thus, $$\frac{\partial^2}{\partial\theta\partial\theta'} \log L(\theta) = \frac{1}{L(\theta)} \frac{\partial^2}{\partial\theta\partial\theta'} L(\theta) - \frac{1}{L^2(\theta)} \frac{\partial}{\partial\theta} L(\theta)\left[\frac{\partial}{\partial\theta} L(\theta)\right]'$$

and $$E\left[-\frac{\partial^2}{\partial\theta\partial\theta'} \log L(\theta)\right] = E\left[\frac{1}{L^2(\theta)} \frac{\partial}{\partial\theta} L(\theta)\left[\frac{\partial}{\partial\theta} L(\theta)\right]'\right]$$

$$-E\left[\frac{1}{L(\theta)} \frac{\partial^2}{\partial\theta\partial\theta'} L(\theta)\right]$$

if both expectations on the right hand side exist.

$$\text{But, } E\left[\frac{1}{L^2(\theta)}\frac{\partial}{\partial\theta}L(\theta)\left[\frac{\partial}{\partial\theta}L(\theta)\right]'\right] = \int_X \frac{1}{L^2(\theta)}\frac{\partial}{\partial\theta}L(\theta)\left[\frac{\partial}{\partial\theta}L(\theta)\right]'L(\theta)dx$$

$$= \int_X \frac{1}{L(\theta)}\frac{\partial}{\partial\theta}L(\theta)\frac{1}{L(\theta)}\left[\frac{\partial}{\partial\theta}L(\theta)\right]'L(\theta)dx$$

$$= \int_X \frac{\partial}{\partial\theta}\text{Log } L(\theta)\left[\frac{\partial}{\partial\theta}\text{Log } L(\theta)\right]'L(\theta)dx$$

$$= E\left[S(\theta)\ S(\theta)'\right]$$

$$= J(\theta)$$

and

$$E\left[\frac{1}{L(\theta)}\frac{\partial^2}{\partial\theta\partial\theta}L(\theta)\right] = \int_X \frac{1}{L(\theta)}\frac{\partial^2}{\partial\theta\partial\theta'}L(\theta).L(\theta)dx$$

$$= \int_X \frac{\partial^2}{\partial\theta\partial\theta'}L(\theta)dx$$

$$= \frac{\partial^2}{\partial\theta\partial\theta'}\int_X L(\theta)dx$$

$$= \frac{\partial^2}{\partial\theta\partial\theta'}.1$$

$$= 0\ .$$

3.5 Iterative Methods for the Solution of the Maximum Likelihood Equation: The Single Parameter Case.

In many instances it is very difficult to find an explicit solution of the Maximum Likelihood equation. In these cases the Maximum Likelihood Estimate is approximated by using iterative techniques. The most popular technique for approximating the Maximum Likelihood Estimate is the method of scoring for parameters. This technique was suggested by Fisher (1925). Several other methods are also used. These include the Newton-Raphson method, the fixed derivative Newton method and the Regula Falsi method and its modifications.

Kale (1961) discusses the Newton-Raphson method, the fixed derivative Newton method and the method of scoring for parameters. He gives sufficient conditions for the convergence of these iterative schemes and applies the three techniques to the solution of a particular problem. Barnett (1966) also discusses these three methods together with the Regula Falsi method, and notes that the Regula Falsi method will be preferred to these three schemes in cases where the likelihood equation has multiple roots. This is so since the Regula Falsi method can be made to scan the range of the likelihood equation to find all the relative maxima.

We now give a brief discussion of a general iterative scheme of which the Newton-Raphson method, the method of scoring

for parameters, the fixed derivative Newton method and the Regula Falsi method are special cases.

The idea is to choose an initial solution t_o, which lies in a neighbourhood of the Maximum Likelihood Estimate, and use an iterative scheme to find the Maximum Likelihood Estimate. t_o is usually chosen to be an unbiased, consistent estimator.

A possible iterative scheme is as follows:

Define $$\phi(\theta) = \theta - \psi(\theta)\frac{d}{d\theta}\ell(\theta),$$

where $\ell(\theta) = \log L(\theta)$.

Consider the iterative scheme

$$t_{n+1} = \phi(t_n) \qquad (3.5.1)$$

$$= t_n - \psi(t_n)\left[\frac{d}{d\theta}\ell(\theta)\right]_{\theta = t_n} . \qquad (3.5.2)$$

Let $e_n = |t_n - \hat{\theta}|$ be the error at the nth iteration, then we choose $\psi(\theta)$ such that $e_{n+1} < e_n$ and $e_n \to 0$ as $n > \infty$.

Householder (1953) has shown that it is sufficient that:

(1) There exists an ε-neighbourhood $N_\varepsilon(\hat{\theta})$ of $\hat{\theta}$, such that if θ_1 and θ_2 are in $N_\varepsilon(\hat{\theta})$, then for some $k \geq 0$, we have

$$|\phi(\theta_1) - \phi(\theta_2)| \leq k < 1,$$

(2) $t_1 \in N_\varepsilon(\theta)$

for (3.5.2) to converge .

The Newton-Raphson method, the method of scoring for parameters, the fixed derivative Newton method and the secant

method* are all special forms of (3.5.1). In the Newton-Raphson method, $\psi(\theta) = \frac{d^2}{d\theta^2}\, \ell(\theta)$.

In the method of scoring for parameters $\psi(\theta) = E_\theta\left[\frac{d^2}{d\theta^2}\, \ell(\theta)\right]$, i.e., $\psi(\theta) = -J(\hat{\theta})$. In the fixed derivative Newton method $\psi(\theta) = k$, where k is a constant such that the conditions for convergence of (3.5.2) hold.

$\psi(\theta) = \left[\frac{d}{d\theta}\ell(\theta) - \left(\frac{d}{d\theta}\ell(\theta)\right)_{\theta = t}\right] \Big/ (\theta - t)$ in the secant method, here t is a consistent estimator of $\hat{\theta}$.

It is well known that the Newton-Raphson method converges quadratically, i.e., $e_{n+1} \propto e_n^2$, whereas the orders of convergence of the other methods lie in the half open interval $[1,2)$. Thus it might seem that the Newton-Raphson method should be preferred to the other methods. However, as pointed out by Barnett (1966) the first three methods listed above all have the same undesirable qualities, and are therefore not recommended in cases where it is suspected that the likelihood equation has multiple roots. Barnett (1966) shows that the Regula Falsi method (and hence its modified forms) works well in many cases where the first three methods fail. For further discussions of iterative techniques the reader is referred to Conte and deBoor (1972) and Hildebrand (1956).

We now use the Newton-Raphson method, the Modified

*This is a modified form of the Regula Falsi method.

Regula Falsi method and the secant method to solve the genetical example given by Fisher (1954) based on Carver's data for two factors in corn; Starchy vs. Sugary and Green vs. White.

We must evaluate the Maximum Likelihood Estimate $\hat{\theta}$ of θ when the probability of belonging to one of the four classes listed below are

Starchy		Sugary	
Green	White	Green	White
$p_1=(2+\theta)/4$	$p_2=(1-\theta)/4$	$p_3=(1-\theta)/4$	$p_4=\theta/4$.

A random sample of size 3839 was taken and the numbers falling into each class are given in the following table:

Starchy		Sugary	
Green	White	Green	White
a = 1997	b = 906	c = 904	d = 32 .

Since the distribution is multinominal, we have

$$L(a,b,c,d|\theta) = k(2 + \theta)^a(1 - \theta)^{b + c}\theta^d ,$$

where k is a constant.

Hence

$$\frac{d}{d\theta}\ell(\theta) = \frac{a}{2+\theta} - \frac{b + c}{1 - \theta} - \frac{d}{\theta} . \qquad (3.5.3)$$

Assuming that $0 < \theta < 1$, it is easy to show that the Maximum

Likelihood Estimate $\hat{\theta}$ can be obtained by solving the following quadratic equation;

$$n\theta^2 + (2b+2c+d-a)\theta - 2d = 0.$$

$\hat{\theta}$ is taken to be the solution which lies in Θ_ℓ and maximizes $\ell(\theta)$. For the above data Fisher (1950) has shown that $\hat{\theta} = .035712$ to six decimal places. Norton (1956) and Kale (1961) use the Newton-Raphson method starting from $t_o = (a-b-c+d)/n = .057046$, and after five iterations obtain $\hat{\theta} = .035712$ and .035713 respectively. We shall start iterating from $t_o = (a+5d-b-c)/2n = .045194$. In the modified Regula Falsi and secant methods we shall use as our two initial estimates $t_{-1} = 4d/n = .033342$ and $t_0 = (a+5d-b-c)/2n = .045194$. It is observed that t_{-1} and t_o are both unbiased and consistent estimators for θ, and that the choice of t_{-1} is justified since

$$\left[\frac{d}{d\theta}\ell(\theta)\right]_{\theta = t_{-1}} = 69.446846 \text{ and } \left[\frac{d}{d\theta}\ell(\theta)\right]_{\theta = t_o} = -211.178969,$$

i.e., $\hat{\theta} \in [t_{-1}, t_o]$. The results for these methods are presented in the table below.

Successive Iterations by the Three Different Methods.

Iterates	I	II	III
t_{-1}	-	.033342	$.03334^2$
t_o	.045194	.045194	.045194
t_1	.033546	.036275	.036275
t_2	.035591	.035379	.035580
t_3	.035712	.035717	.035714
t_4	-	.035712	.035712

I denotes the Newton-Raphson Method.

II denotes the modified Regula Falsi Method.

III denotes the secant method.

It is observed that the secant method performs just as well as the Newton-Raphson method. We also observe that the Newton-Raphson method with our choice of t_o takes one less iteration than with the t_o used by Kale (1961) and Norton (1956).

We note here that the Newton-Raphson method, the method of scoring for parameters, and the fixed derivative Newton method can be applied to the multiparameter case.

3.6 Properties of MLE's Obtained from Dependent Observetions

Under certain regularity conditions (mentioned earlier) Cramér (1946) proved that the MLE obtained from independent observations is:

(i) weakly consistent, and

(ii) asymptotically efficient and normally distributed. These results were later extended by other authors, some of whom relaxed a few of Cramér's condition, while others imposed additional conditions on the likelihood equation.

In recent times Silvey (1961), Bar-Shalom (1970) and Bhat (1973) considered the asymptotic properties of MLE's obtained from dependent observations. Bar Shalom (1970) imposed two conditions on the likelihood equation in addition to those stated by Cramér. He then used a version the the weak law of large numbers for dependent random variables to prove the convergence of b_o and b_1 (defined below). Bhat (1973) used the central limit theorem for Martingale sequences to prove the convergence of b_o and b_1.

We now list the Regularity conditions as given by Bar-Shalom (1970).

Notation and Listing of Regularity Conditions.

Let the set of (possibly dependent) observations be $Z^n = (z_1, z_2, \ldots, z_n)$, where z_i, $i = 1, \ldots, n$, are real valued random variables with joint probability density function (p.d.f) With respect to a σ-finite (i.e., finite with respect to the σ-algebra generated by Z^n) measure μ_n, given by

$$P(z_1, z_2, \ldots, z_n|\theta) = P(Z^n|\theta) \qquad (3.6.1)$$

where θ is a real valued constant with unknown, but true

value $\theta_o \in \Theta$, with Θ an open subset of $|R^1$. The Borel measurable function $\hat{\theta}_n(Z^n)$ obtained by maximization of the likelihood function

$$L_n(\theta) = P(Z^n|\theta) = \prod_{i=1}^{n} p_i(\theta) ,$$

where $p_i(\theta) = L_i(\theta)/L_{i-1}(\theta)$ is used as an estimator for θ.

We now list the following regularity conditions which are needed to prove the asymptotic properties of the MLE.

RC1. For almost all Z^k, $\frac{\partial^i}{\partial\theta^i} \log p_k(\theta)$, i=1,2,3, exist for all $\theta \in \Theta$.

RC2. $E\left[\frac{\partial}{\partial\theta} \log p_k(\theta)\right] = 0 .$

RC3. $J_k(\theta) = E\left[\left[\frac{\partial}{\partial\theta} \log p_k(\theta)\right]^2\right]$
$= \int_{|R^k} \left[\frac{\partial}{\partial\theta} \log p_k(\theta)\right]^2 L_k(\theta) \prod_{i=1}^{k} \mu(dy_i) \leq C_1 < \infty ,$

where C_1 is independent of θ, and $J_k(\theta)$ is Fisher's information measure (see for example Rao (1973)).

RA4. $E\left[\frac{\partial^2}{\partial\theta^2} \log p_k(\theta)\right] = - J_k(\theta_o) .$

RC5. There exists a function, $B_k(Z^k)$, measurable with respect to the product measure $\prod_{i=1}^{k} \mu(dy_i)$,

such that $\left|\frac{\partial^3}{\partial\theta^3} \log p_k(\theta)\right| < B_k(Z^k)$ for all θ ,

and $B_k(Z^k)$ is finite almost everywhere, i.e., there exists a constant, $N < \infty$, independent of θ and k,

such that for all $\varepsilon > 0$, $P(B_k(Z^k) > N) < \varepsilon$.

RC6. $$E\left[\frac{\partial}{\partial\theta} \log p_i(\theta) \; \frac{\partial}{\partial\theta'} \log p_j(\theta)\right] = 0, \text{ for all } i \neq j.$$

A milder version of RC6 is;

RC7. $$\lim_{|i-j| \to \infty} E\left[\frac{\partial}{\partial\theta} \log p_i(\theta) \; \frac{\partial}{\partial\theta'} \log p_j(\theta)\right] = 0 .$$

RC8. $$\text{Var}\left[\frac{\partial^2}{\partial\theta^2} \log p_i(\theta)\right] \leq C_2 < \infty$$

where C_2 is a constant independent of i, and

$$\lim_{|i-j| \to \infty} \text{Cov}\left[\frac{\partial^2}{\partial\theta^2} \log p_i(\theta) \; \frac{\partial^2}{\partial\theta^2} \log p_j(\theta)\right] = 0 .$$

Note: In the above regularity conditions all the expectations are evaluated at, $\theta = \theta_o$, the true value. Conditions RC1 to RC5 are similar to the conditions RA1 to RA5 given by Cramér (1946). Conditions RC6 to RC8 are needed because of the possible dependence among the observations.

We are now in a position to state and prove the following theorem.

Theorem: Under regularity conditions RC1 to RC8, the Maximum Likelihood Estimator $\hat{\theta}_n(Z^n)$ is weakly consistent, i.e.,

$$\hat{\theta}_n(Z^n) \xrightarrow{p} \theta_o,$$

(where $\xrightarrow{p}$ denotes convergence in probability) and asymptotically efficient, i.e., for sufficiently large n the mean square of $\hat{\theta}_n(Z^n)$ is equal to the reciprocal of the total information, i.e.,

$$E\left[(\hat{\theta}_n-\theta_0)^2\right] = \left[\sum_{i=1}^{n} J_i(\theta_0)\right]^{-1} .$$

Proof: Following Cramér, by conditions RC1 to RC6

$$\frac{\partial}{\partial\theta}\log p_k(\theta) = \left[\frac{\partial}{\partial\theta}\log p_k(\theta)\right]_{\theta=\theta_0} + (\theta-\theta_0)\left[\frac{\partial^2}{\partial\theta^2}\log p_k(\theta)\right]_{\theta=\theta_0} + \frac{\lambda}{2}(\theta,\theta_0)\cdot(\theta-\theta_0)^2 B_k(Z^k) , \qquad (3.6.2)$$

where B_k is defined above, and $|\lambda| < 1$.

Thus the likelihood equation (3.6.2) may, after multiplication by $\frac{1}{n}$ and summing, be written as

$$\frac{1}{n}\frac{\partial}{\partial\theta}L_n(\theta) = b_0 + (\theta-\theta_0)b_1 + \frac{\lambda}{2}(\theta-\theta_0)^2 b_2 = 0 ,$$

where
$$b_o = \frac{1}{n}\sum_{i=1}^{n}\left[\frac{\partial}{\partial\theta}\log p_i(\theta)\right]_{\theta=\theta_0}$$
$$b_1 = \frac{1}{n}\sum_{i=1}^{n}\left[\frac{\partial^2}{\partial\theta^2}\log p_i(\theta)\right]_{\theta=\theta_0}$$
$$b_2 = \frac{1}{n}\sum_{i=1}^{n} B_i(Z^i).$$

We now use a version of the weak law of large numbers for dependent random variables (Parzen (1969), p. 418) to prove the convergence of b_o and b_1. This law states that if for all i the random variable x_i has finite mean and bounded variance and $\lim_{n\to\infty}\frac{1}{n}\sum_{i=1}^{n} E(x_i) < \infty$, then a sufficient condition for $\frac{1}{n}\sum_{i=1}^{n} x_i - \frac{1}{n}\sum_{i=1}^{n} E(x_i) \xrightarrow{p} 0$, is $\lim_{n\to\infty}\frac{1}{n}\sum_{i=1} \mathrm{Cov}(x_i,x_j) = 0$.

Using this theorem, it follows from RC2, RC3 and RC7 that, $b_0 \xrightarrow{p} 0$,

and from RC4 and RC8 that, $b_1 \xrightarrow{p} - J(\theta_0)$, where

$J(\theta_0) = \lim_{n \to \infty} \frac{1}{n} \sum_{i=1}^{n} J_i(\theta_0)$ exists by virtue of RC3. From RC5, it follows that for all $\varepsilon > 0$, there exists N such that

$$P(b_2 \geq N) < \varepsilon .$$

We now use these results to show that the estimator $\hat{\theta}_n(Z^n)$ is weakly consistent.

Let δ and ε be given arbitrarily small positive numbers. For sufficiently large $n(n > n_o(\delta,\varepsilon)$, say) we have

$$\begin{aligned} P_0 &= P(|b_o| \geq \delta^2) < \tfrac{\varepsilon}{3} \\ P_1 &= P(b_1 \geq -\tfrac{1}{2} J(\theta_0)) < \tfrac{\varepsilon}{3} \\ P_2 &= P(|b_2| \geq 2N) < \tfrac{\varepsilon}{3} . \end{aligned} \qquad (3.6.3)$$

If we let S denote the set of all points Z^n such that, $|b_0| < \delta^2, b_1 < -J(\theta_0)/2$ and $|b_2| < 2N$ are all satisfied, then the complementary set S' will consist of all points Z^n, such that at least one of the inequalities is not satisfied. Therefore, by an elementary law of probability, we have that

$$P(S') \leq p_0 + p_1 + p_2 < \varepsilon .$$

Hence, $P(S) > 1-\varepsilon$, and thus $P(Z^n \in S) > 1-\varepsilon$ whenever $n > n_o(\delta,\varepsilon)$.

For $\theta = \theta_o \pm \delta$, the right hand side of equation (3.6.2) becomes

$$b_o \pm b_1\delta + \frac{\lambda}{2} b_2\delta^2 .$$

Hence for every point $Z^n \in S$,

$$\left|b_o + \frac{\lambda}{2} b_2\delta^2\right| < (N+1)\delta^2 ,$$

and $$b_1 \delta < J(\theta_0)\delta/2 .$$

Now if $\delta < J(\theta_o)/2(N+1)$, the sign of the right hand side of (3.6.2) will be determined by the sign of $\pm b_1\delta$, for $\theta = \theta_0 \pm \delta$, hence $\frac{\partial}{\partial\theta} L_n(\theta) > 0$ for $\theta = \theta_o - \delta$ and $\frac{\partial}{\partial\theta} L_n(\theta) < 0$ for $\theta = \theta_o + \delta$. Furthermore, by RC1, the function $\frac{\partial}{\partial\theta} L_n(\theta)$ is a.e.(Z^n)* a continuous function of $\theta \in \Theta$. Therefore, for sufficiently small δ and ε, the likelihood equation has a root (in the open interval $(\theta_o - \delta, \theta_o + \delta)$) with probability larger than $1-\varepsilon$, for $n > n_o(\delta,\varepsilon)$.

The proof of asymptotic efficiency follows Bar-Shalom (1970).

Let $\hat{\theta}_n$ be a root of the likelihood equation (3.6.2) in the interval $(\theta_o-\delta, \theta_o+\delta)$. We can arrange equation (3.6.2) as follows

*a.e. denotes almost everywhere

$$\hat{\theta}_n - \theta_o = -\frac{b_o}{J(\theta_o)} \Big/ \left[\frac{b_1}{J(\theta_o)} + \frac{\lambda}{2J(\theta_0)}(\theta_n - \theta_o) b_2\right] . \qquad (3.6.4)$$

Since for $n > n_o(\delta, \varepsilon)$, $|\hat{\theta}_n - \theta_o| < \delta$ with probability exceeding $1-\varepsilon$. Using (3.6.3) we can show that the denominator of (3.6.4) converges to 1. From RC2, it follows that the numerator of (3.6.4) has mean value zero and variance given by

$$E\left[(b_o/J(\theta_o))^2\right] = \frac{1}{(J(\theta_o))^2} \; E\left[\left(\frac{1}{n}\sum_{i=1}^{n} \frac{\partial}{\partial\theta} \log P_i(\theta)\right)^2\right]$$

$$= \left[nJ(\theta_o)\right]^{-2} \; E\left[\left(\sum_{i=1} \frac{\partial}{\partial\theta} \log p_i(\theta)\right)^2\right],$$

which, for $n > n_o(\delta, \varepsilon)$, is equal to the variance of $\hat{\theta}_n$. Hence the total information for the parameter θ from n dependent observations is given by

$$J^n(\theta_o) = (nJ(\theta_o))^2 \Big/ \left[E\left[\sum_{i=1}^{n} \frac{\partial}{\partial\theta} \log p_i(\theta)\right]^2\right]_{\theta = \theta_o} . \qquad (3.6.5)$$

It is easily seen from the additivity property of information that the total information is equal to the sum of the information contained in each observation, if and only if condition RC6 is satisfied.

The denominator of (3.6.5) can be written as

$$E\left[\sum_{i=1}^{n} \left[\frac{\partial}{\partial\theta} \log p_i(\theta)\right]^2 + 2 \sum_{i<j}^{n} \frac{\partial}{\partial\theta} \log p_i(\theta) \frac{\partial}{\partial\theta} \log p_j(\theta)\right], \text{ which}$$

on using RC6, with $n > n_o(\delta, \varepsilon)$, becomes

$$\sum_{i=1}^{n} E\left[\frac{\partial}{\partial\theta} \log p_i(\theta)\right]^2 = \sum_{i=1}^{n} J_i(\theta_0)$$

$$= nJ(\theta_0).$$

Thus, for $n > n_o(\delta, \varepsilon)$, we may rewrite equation (3.6.5) as

$J^n(\theta_0) = n\, J(\theta_o)$,

and so

$$E\left[(\hat{\theta}_n - 0_o)^2\right] = nJ(\theta_o)$$

$$= \sum_{i=1}^{n} J(\theta_o).$$

Q.E.D.

CHAPTER IV

LEAST SQUARES

4.1 Introduction. In most scientific investigations it is frequently necessary to determine the values of certain parameters by means of actual measurements. The observations may be made directly on the values of the unknown quantities (dependent variables) or on certain functions of these quantities. In the latter case, the values of the required quantities must be computed from the observed values of the functions. In order to obtain reasonably accurate results, the observations are usually repeated, either in the same way under the same conditions, or in different ways under varying conditions.

Under these circumstances it will be found that different measurements of the same quantity usually give discordant results, the amount of discrepancies varying with the mode of observation. The question, we are faced with now, is how to determine from these discordant results the true values of the required parameters. Because of these discrepancies in the observations, we cannot expect to obtain our parameters with absolute accuracy. All that we can hope for is to obtain those values which are rendered most probable when all the observations are taken into account.

We illustrate the above difficulties with a concrete example. We are required to determine the coefficient of linear

expansion of a certain metal rod, based on measurements of the length of the rod at various temperatures.

Temperature in ^{o}C	Observed Length in m.m.
20	1000.22
40	1000.65
50	1000.90
60	1001.05

Let c denote the required coefficient of linear expansion and ℓ_o the length of the rod at $0^{o}C$. The length ℓ of the rod at any other temperature t may be represented by the equation:

$$\ell = \ell_o + ct \qquad (4.1.1)$$

Using this model the data can be transformed into the following set of equations:

$$\ell_o + 20c = 1000.22 \qquad (4.1.2)$$

$$\ell_o + 40c = 1000.65 \qquad (4.1.3)$$

$$\ell_o + 50c = 1000.90 \qquad (4.1.4)$$

$$\ell_o + 60c = 1001.05 \; . \qquad (4.1.5)$$

We can use any two of the above equations to determine the values of ℓ_0 and c, but these values will depend on the particular pair of equations used, and in general will be different for each pair. For example, we may find from the following pairs of equations, the different values for ℓ_o and c:

Equations	ℓ_0(m.m.)	c(m.m./^{0}C)
(4.1.2) and (4.1.3)	999.790	.02150
(4.1.2) and (4.1.4)	999.767	.02267
(4.1.3) and (4.1.4)	999.805	.02025
(4.1.4) and (4.1.5)	1000.150	.01500
etc.	etc.	etc.

We are faced with the problem of choosing the best (in the sense of fitting the observations most closely) pair of results. However, in such a situation, we would be disregarding the bulk of the observations thereby losing the majority of the information available. In order to make use of all the data, we choose as our estimates of the parameters those values which make the sum of squared deviations of the observed values from the predicted values a minimum. The name given to this process is the method of Least Squares.

The theory of Least Squares was first discussed by Legendre (1805) and Gauss (1805) who used it as a tool of estimation. Markov and many other authors have since made significant contributions to the general theory.

The underlying assumption in the models of Gauss and Markov is that the error covariance matrix is the identity matrix. Aitken (1934) proposed a model in which the error covariance matrix was a symmetric positive definite matrix, not necessarily the identity. This model was later shown to be equivalent to

the model of Gauss and Markov. Rao (1972) developed a unified theory of Least Squares by which any problem involving regression coefficients can be analysed.

The theory of Least Squares, unlike the method of Maximum Likelihood, has no general optimum properties to recommend it in the finite sample case. However, there exists an important class of situations in which it does provide unbiased estimators (which are linear functions of the observations) with minimum variance. We shall also see that, contrary to Kendall and Stuart Vol. II (1960), Least Squares estimators do have asymptotic properties akin to those of Maximum Likelihood estimators.

The solution of the Least Squares Equation is discussed in many standard texts. We shall therefore not discuss its solution explicitly. Instead, we shall discuss properties of Least Squares estimators under different conditions.

4.2 Asymptotic Properties of Least Squares

Eicker (1962) considers the linear regression model

$$y = X_n\beta + \varepsilon \qquad (4.2.1)$$

where the subscript n denotes the dependence of the matrix X on the sample size. It is assumed that the components of the error vector ε are either (a) uncorrelated or (b) independent with $E[\varepsilon] = 0$ and $0 < \text{Var}(\varepsilon) < \infty$. The distributions of the components ε_k of ε are not assumed to be known, nor are

they assumed to be identical. However, it is assumed that the ε_k are elements of a certain set F of distribution functions. With these assumptions Eicker (1962) gives conditions on the set F and on the matrix X_n such that the least squares estimators of the parameters $\beta_1, \ldots, \beta_q$ are consistent in case (a) or asymptotically normal in case (b).

Cox and Hinkley (1967) consider a similar model. However, they assume that the first column of the matrix X_n consists of all ones and that the components ε_k of the error vector ε are independent and identically distributed with known probability density $f(\varepsilon,\lambda)$, with zero mean, where λ is an unknown parameter which gives the spread and possibly also the shape of the distribution of f. They then give a condition under which the least squares estimators of $\beta_1,\beta_2,\ldots,\beta_q$ are asymptotically normal, and then use the fact that the maximum likelihood estimators of $\beta_1,\ldots,\beta q$ are asymptotically normal (see Chapter 3), to show the asymptotic efficiency of the Least Squares estimators.

We shall follow Cox and Hinkley (1967) for the discussion of asymptotic efficiency, and Eicker (1962) for the discussion of consistency.

As noted by Cox and Hinkley (1967), the general mean, although included in the linear regression model, is seldom a parameter of primary interest. We therefore can, without loss of generality, assume that the first column of X_n consist

of all ones and that the other column sums of X_n are zero. We can also take the parameters $\beta_2,\ldots,\beta_q$ to be orthogonal to β_1, the general mean.

Since the distributions of the errors are known, we can use the method of Maximum Likelihood to obtain the least squares estimators of the parameters $\beta_1,\ldots,\beta_q$. The likelihood function L is given by

$$L = \prod_{i=1}^{n} f(\varepsilon_i,\lambda) \quad . \tag{4.2.2}$$

Thus the log-likelihood is given by

$$\ell = \sum_{i=1}^{n} \log f(\varepsilon_i,\lambda) = \sum_{i=1}^{n} g(\varepsilon_i,\lambda), \quad (4.2.3), \text{ say .}$$

We shall assume that the range of regularity of ℓ does not depend on the parameters $\beta_1,\beta_2,\ldots,\beta_q$.

Let $j, k = 1,\ldots, q$. Then for fixed λ

$$\frac{\partial \ell}{\partial \beta_k} = \sum_{i=1}^{n} g'(\varepsilon_i,\lambda)\frac{\partial \varepsilon}{\partial \beta_k}_k, \tag{4.2.4}$$

and

$$\frac{\partial^2 \ell}{\partial \beta_j \partial \beta_k} = \sum_{i=1}^{n} g'(\varepsilon_i,\lambda)\, \frac{\partial \varepsilon_j}{\partial \beta_j}\frac{\partial \varepsilon_k}{\partial \beta_k}$$

$$= \sum_{i=1}^{n} g''(\varepsilon_i,\lambda)\, X_{ij}X_{ik} \;, \tag{4.2.5}$$

since $\varepsilon_i = y_i - \sum_{k=1}^{q} X_{ik}\beta_k$.

Thus $$E\left[\frac{-\partial^2 \ell}{\partial \beta_j \partial \beta_k}\right] = X'_j\, X_k\, A_{\varepsilon_i} \;, \tag{4.2.6}$$

Where X_j is the j^{th} column of X_n and

$$A_{\varepsilon_i} = \int_{|R^1} f(\varepsilon_i,\lambda) \frac{\partial^2}{\partial \varepsilon_i^2} \log f(\varepsilon_i,\lambda) d\varepsilon_i \ .$$

If λ_1 is a component of the parameter λ, we have

$$\frac{\partial^2 \ell}{\partial \beta_j \partial \lambda_1} = - \sum_{i=1}^{n} \frac{\partial}{\partial \lambda_1} g'(\varepsilon_i,\lambda) X_{ij} \ , \qquad (4.2.7)$$

and since $\sum_{i=1}^{n} X_{ij} = 0$, $j = 2,\ldots,q$ we have

$$E\left[\frac{\partial^2 \ell}{\partial \beta_j \partial \lambda_1}\right] = 0 \ .$$

Similarly,

$$E\left[\frac{\partial^2 \ell}{\partial \beta_j \partial \beta_1}\right] = 0 \ , \ j = 2,\ldots,q \ .$$

Thus the information matrix for $\begin{bmatrix} \beta \\ \lambda \end{bmatrix}$ is

$$J\begin{bmatrix} \beta \\ \lambda \end{bmatrix} = \begin{bmatrix} E & 0 \\ 0 & H \end{bmatrix} \ ,$$ where E is the information matrix for the parameters β_1 and λ, and H the information matrix for the parameters $\beta_2,\ldots,\beta_q$, with

$$H = \left[X_j' X_k A_{\varepsilon_i}\right]_{h,\ell} \qquad h,\ell = 2,\ldots,q \ .$$

By the orthogonality condition, the covariance matrix of the Least Squares estimators of $\beta_2,\ldots,\beta_q$ is

$$\left[\left[X_j' X_k\right]_{j,k}\right]^{-1} \quad \text{Var}\ (\varepsilon) \ .$$

Thus using the fact that the maximum likelihood estimators of $\beta_2,\ldots,\beta_q$, have covariance matrix

$$\left[\left(X_j' X_k\right)_{j,k}\right]^{-1}$$

we have that the asymptotic efficiency is $\left[A_{\varepsilon_i} \text{Var}(\varepsilon_i)\right]^{-1}$

which is independent of the design matrix X_n.

Asymptotic Efficiency

In the model considered by Eicker (1962) the design matrix X_n (n x q) is assumed to be of full rank $q \leq n$.

The normal equations obtained by minimizing $\varepsilon'\varepsilon$ are

$$X_n' y = X_n' X_n \tilde{\beta} \tag{4.2.8}$$

where $\tilde{\beta} = \hat{\beta}_n$ is the vector of least squares estimators of β. Let $H_n = X_n' X_n$; by virtue of the rank assumption on X_n, H_n^{-1} exists, and so

$$\tilde{\beta} - \beta = H_n^{-1} X_n' \varepsilon \tag{4.2.9}$$

$$\mathrm{Var}(\tilde{\beta} - \beta) = H_n^{-1} X_n' V X_n H_n^{-1}, \tag{4.2.10}$$

where $V = \mathrm{var}(\varepsilon)$.

We now state and prove a theorem which gives necessary and sufficient conditions for the consistency of least squares estimators.

Theorem: A necessary and sufficient condition for the least squares estimator $\tilde{\beta}$ to estimate β consistently on F, is that $Ch_{min} H_n \to \infty$, where $Ch_{min} H_n$ is the smallest characteristic root of H_n.

Proof:

Sufficiency: Since $E\left[H_n^{-1} X_n' \varepsilon\right] = 0$, the estimator $\tilde{\beta}$ is unbiased. Now, the variance of each component of the vector $H^{-1} X_n' \varepsilon$ tends to zero if and only if $E\left[\varepsilon' X_n H_n^{-2} X_n' \varepsilon\right]$ tends to

zero.

Since $E\left[\varepsilon_k^2\right]$ is assumed constant (i.e., independent of k and n), we have

$$E\left[\varepsilon' X_n H_n^{-2} X_n' \varepsilon\right] = 0\ (tr(X_n H_n^{-2} X_n'))$$
$$= 0\ (tr(H_n^{-1}))$$
$$= 0\ (1/Ch_{min} H_n)\ , \qquad (4.2.11$$

where $0(n)$ = constant x n.

Therefore, since $Ch_{min} H_n \to \infty$, we have

$$\tilde{\beta} \xrightarrow{p} \beta \quad \text{on F.}$$

Necessity: If we choose all $\varepsilon_k \sim N(0,\sigma^2) \in F$

we see that $$\tilde{\beta} - \beta = H_n^{-1} X_n' \varepsilon \sim N_q\left[0, \sigma^2 H_n^{-1}\right].$$

Hence the variance of the ith component of $\tilde{\beta} - \beta$ is given by $Var\ (\tilde{\beta} - \beta)_i = \sigma^2 (H_n^{-1})_{ii}$, which because of consistency must tend to zero for every $i = 1,\ldots,q$.

Therefore $$tr\ H_n^{-1} = \sum_{i=1}^{q} Ch_i H_n^{-1}$$

must tend to zero and so $Ch_{min} H_n$ must tend to infinity.

Q.E.D.

4.3 Robustness of the Method of Least Squares.

A statistical procedure is said to be robust if it is insensitive to departures from the assumptions which underlie it.

In the case of the general linear model, if the error vector has mean zero and dispersion matrix $\sigma^2 I_n$, where σ^2 is an unknown constant, then the estimates of the parameters and their variances will remain valid even if the error vector is not normally distributed; i.e., the method of least squares is robust to non-normality. Box and Watson (1962) have studied the robustness of tests on the regression coefficients with respect to the non-normality of the error vector. They mention that for a particular design matrix X(nxp), the variance mean-square ratio R_m (obtained in comparison of means), and R_v (obtained in comparison of variances) have distributions which may be approximated by F distributions with modified degrees of freedom. Box and Andersen (1955) showed that the degrees of freedom for the distribution of R_m are $\upsilon_1 = \delta_m p$ and $\upsilon_2 = \delta_m (n-p-1)$ and for R_v are $\upsilon_1 = \delta_v p$ and $\upsilon_2 = \delta_v (n-p-1)$, where for mild departures from normality and for moderate numbers of observations,

$$\delta_m^{-1} \simeq 1 - (1/n)\,\beta_1 \quad \text{and} \quad \delta_v^{-1} \simeq 1 + \tfrac{1}{2}\,\beta_2 \text{ , where}$$

β_1 and β_2 are measures of kurtosis, which have values zero when the distribution is normal. It is therefore seen that R_m will be insensitive to non-normality because the corrective

factor is of order n^{-1}.

When the assumption of uncorrelated homoscedastic errors is not satisfied, the least squares estimate of γ,

$$\hat{\gamma} = (X'X)^{-1}X'Y$$

is still unbiased; however it no longer has the property of minimum variance. Thus, heteroscedasticity and correlation of the errors still leaves the least squares estimator unbiased. However, the efficiency of such estimators is reduced. It should be pointed out here that the method of least squares is not robust with respect to biased observations.

4.4 A Different Computational Technique

There are many sophisticated techniques available for solving the normal equations. These techniques include, singular value decomposition and Householder transformation. We present here an algorithm which finds least squares estimates by minimization of the error sum of squares, rather than by solving the normal equations. The method involves the use of an iterative scheme, and it can also be used to find Ridge estimates.

Consider the general Gauss-Markov model

$$Y = X\gamma + u$$

where X is an $n \times q$ matrix with $q \leq n$ for which $r = r(X) \leq q$, $E[u] = 0$ and $Var[u] = \sigma^2 I$.

The least squares estimator $\hat{\gamma}$ of γ is obtained by mini-

mizing

$$||y-X\gamma||_2^2 = u'u. \tag{4.4.1}$$

Now, $$||y-X\gamma||_2^2 = y'y-2\gamma'X'y + \gamma'X'X\gamma. \tag{4.4.2}$$

So consider $$||y-X\gamma||_2^2 + \alpha||X||_2^2. \tag{4.4.3}$$

It was shown in Chapter I (1.31) that as α tends to zero the least squares estimate $\hat{\gamma}(\alpha)$ of (4.4.3) converges to the least squares estimate $\hat{\gamma}$ of (4.4.2). We shall use this fact and an iterative technique called the Davidon-Fletcher-Powell Method (DFP), (see Walsh (1975)), to solve (4.4.2).

Using (4.4.3), we have,

$$\begin{aligned} ||y - X\gamma||_2^2 + \alpha||\gamma||_2^2 &= \gamma' X'X\gamma + \alpha\,\gamma'\gamma - 2\gamma'X'y + y'y \\ &= \gamma'\left[X'X + \alpha I\right]\gamma - 2\gamma'X'y \\ &\quad + y'y . \end{aligned} \tag{4.4.4}$$

Davidon considers quadratic functions of the form

$$\tfrac{1}{2}\gamma'G(\alpha)\gamma + \gamma'b + c, \tag{4.4.5}$$

Where $G(\alpha)$ is symmetric and positive definite;

here $G(\alpha) = X'X + \alpha I > 0$ for all $\alpha > 0$.

(4.4.3) is minimized when

$$\frac{\partial}{\partial\gamma} f(\gamma,\alpha) = 2G(\alpha)\alpha - 2X'y = 0, \tag{4.4.6}$$

which gives $\hat{\gamma}(\alpha) = \left[G(\alpha)\right]^{-1}X'y$. However, for $n > 3$, it is usually difficult to invert $G(\alpha)$, and so we use other methods to find $\hat{\gamma}(\alpha)$ explicitly.

Notation:

$$\text{Let } g_k = g(\gamma_k) = \frac{\partial}{\partial \gamma} f(\gamma,\alpha)\Big|_{\gamma = \gamma_k}$$

The Method

1/ Set $d_k = -H_k g_k$,

with $H_1 = I$. Then d_k is in the direction of search from the current point γ_k.

2/ Find λ_k^* which minimizes $f(\gamma_k + \lambda_k d_k)$.

3/ Set $\sigma_k = \lambda_k^* d_k$.

4/ Set $\gamma_{k+1} = \gamma_k + \sigma_k$.

5/ Evaluate $f(\gamma_{k+1})$ and g_{k+1} , noting that g_{k+1} is orthogonal to σ_k, i.e. $g'_{k+1}\,\sigma_k = 0$.

6/ Set $\omega_k = g_{k+1} - g_k$.

7/ Set (i) $A_k = \sigma_k\,\sigma_k' / \sigma_k'\,\omega_k$,

(ii) $B_k = -H_k\,\omega_k\,\omega_k'\,H_k \,/\, \omega_k'\,H_k\,\omega_k$,

(iii) $H_{k+1} = H_k + A_k + B_k$.

8/ Set $k = k + 1$.

9/ Check $||\sigma_k||$ or $||d_k||$, if either satisfies the desired stopping criterion, then stop, else return to Step 1 and repeat the process.

For an $n \times n$ system (i.e. n variables), the scheme will converge in at most n steps. Fletcher and Powell recommend that we perform one extra iteration after the apparent minimum is attained. This will help

to avoid false minima.

4.4a A Further Application of the Davidon-Fletcher-Powell Method.

We noted in Chapter two that a desirable property of estimators is that of unbiasedness. However, as shown by Hoerl, Kennard and others, in many problems in multiple regression, the Least Squares estimates have a high probability of being unsatisfactory, or even incorrect, especially when the matrix $X' X$ is singular. Hoerl and Kennard (1970) have shown that the addition of a scalar matrix kI_p, $k \geq 0$, to $X' X$ usually corrects this problem, and for a particular choice of k, we obtain a point estimate of the regression parameter which has a smaller mean square error than than the Least Squares Estimate. This is a particular case of biased estimation and is called Ridge Regression. We note here that in view of (1.31), the Davidon-Fletcher-Powell method can be modified so as to produce Ridge estimates and calculate the mean square error.

Example (Goldberger):

Consider the following macro-economic production function

$$y_t = \beta_1 X_{t_1} + \beta_2 X_{t_2} + \beta_3 X_{t_3} + \beta_4 X_{t_4} + \varepsilon_t,$$

where y_t is the real gross national product in billions of dollars. $X_1 = 1$, X_2 is labour inputs in millions of man-years, X_3 is real capital in billions of dollars, measured from an arbitrary origin and X_4 is the time in years measured from

1929 ≡ 1, as the base year. The sample consists of 23 annual observations for the United States from 1929 to 1941 and 1946 to 1955.

The observations are as follows.

X_2	X_3	X_4	y
47	54	1	142
43	59	2	127
39	57	3	113
34	48	4	98
34	36	5	94
36	24	6	102
38	19	7	116
41	18	8	128
42	22	9	140
37	24	10	131
40	23	11	143
42	27	12	157
47	36	13	182
51	9	18	209
53	25	19	214
53	39	20	225
50	51	21	221
52	62	22	243
54	75	23	257
54	94	24	265

55	108	25	276
52	118	26	271
54	124	27	291

This data was analysed using the Davidon-Fletcher-Powell method, and convergence was achieved after four iterations.

The solution is

$$y = -61.6196 + 3.8116\ X_2 + .3121\ X_3 + 3.8474\ X_4 \ .$$

The solution given by Goldberger is

$$y = -\ 61.728 + 3.8191\ X_2 + .3219\ X_3 + 3.7862\ X_4 \ .$$

Golberger's computation was done using six decimal digit accuracy whereas our computation was done using 15 decimal accuracy and is therefore more accurate.

The error sums of squares is 163827.8 .

The computations were made on the IBM 360/370 computer of the McGill Computing Center.

Results for Goldberger's Example

Iterate	β_1	β_2	β_3	β_4	F	$\left\|\left\|\frac{\partial F}{\partial \beta}\right\|\right\|$	σ
1	-50.00000	5.00000	0.0	5.00000	133682.0	345052.00	-
2	-50.01198	4.44023	-0.60235	4.81897	940607.1	33455.75	.000004
3	-50.03816	3.36594	0.26694	4.59982	-300116.2	1145.11	.000065
4	-50.02924	3.48767	0.31627	4.04777	-245180.0	5.91	2.092224
5	-61.61956	3.81155	0.31214	3.84736	163827.8	$< 10^{-10}$	.7550133

4.5 Estimability of Linear Contrast in the General Linear Model (GLM)

The concept of estimability, that is, the existence of unbiased linear estimators for linear contrasts in the general linear model with less than full rank, was introduced by Bose (1944). This concept was developed because the conventional methods, which relied on the inversion of a matrix could not be applied when the matrix is singular. The concept was later discussed by many other authors: Roy and Roy (1959, 1961) who gave a condition for estimability; Searle (1965) who seems to be the first author to use generalized inverses to characterize estimable functions; Miliken (1971) who gives two equivalent conditions for estimability which are specific forms of the condition given by Roy and Roy (1961) and Rao (1972) who developed a unified theory of Least Squares.

In this section we shall discuss some of the many results stated by these and other writers. We shall also prove some theorems on the various characterizations of estimability.

We consider the general linear model

$$y = X\beta + \varepsilon$$

where y is an $n \times 1$ random observation vector, X is an $n \times p$ matrix of known constants, of rank q $(q < p)$, β is a $p \times 1$ vector of unknown parameters defined in R_p, and ε is an unobservable random vector with mean 0 and covariance matrix $\sigma^2 V$ where σ^2 is positive and unknown, and V is a non-negative

definite matrix. When V is positive definite, we can re-parameterize the model so that the resulting error vector has covariance matrix $\sigma^2 I$, we shall therefore restrict our discussion to the latter case.

A considerable amount of research has been recently done on the general linear model in which the error vector has a singular covariance matrix. The most important discussions on this model are due to Rao (1971, 1972, 1973a, 1973b, 1973c, 1974, 1975). This model was also studied by several other authors, including Zyskind (1967), Zyskind and Martin (1969), and Alalouf (1975), who,noting that the unified theory developed by Rao (1971, 1972, 1973a) failed to provide an explicit algebraic criterion for a function $A'y$ to be unbiased for the linear contrast $L'\beta$, developed a theory similar to Rao's unified theory of Least Squares to handle this problem. We shall present some of the main ideas from Rao (1971, 1972, 1973a, 1973b, 1973c) and Alalouf (1975).

We now state and prove some theorems relevant to the forthcoming discussion.

Definition: The linear contrast $L'\beta$ is said to be estimable, whenever there exists a linear unbiased estimator $A'y$, such that;

$$E\left[A'y\right] = L'\beta = A'X\beta \qquad (4.5.1)$$

for all values of β. Then $L' = A'X$, if and only if $L'\beta$ is estimable.

Theorem (Roy and Roy (1961)):

The linear contrast $L'\beta$ is estimable if and only if

$$r(XT) = r(X) - r(L), \tag{4.5.2}$$

where $T \in \mathcal{N}(L)^*$.

Proof: Using (1.26), we have,

$$r\begin{bmatrix} X \\ L' \end{bmatrix} = r(L') + r\left[X(I - (L')^{-}L')\right] . \tag{4.5.3}$$

The quantity on the right of equation (4.5.3) is equal to $r(X)$ if and only if

$$r\left[X(I - (L')^{-}L')\right] = r(X) - r(L),$$

but $I - (L')^{-}L' \in \mathcal{N}(L')$ so $I - (L')^{-}L'$ is a possible choice of T. We now show that XT is invariant under any choice of T.

Let $T = I-(L')^{-}L' = UV'$ (a full rank decomposition) where U and V are $q \times t$ matrices each of rank

$$t = r(T) = q - r(L) .$$

Then

$$\begin{aligned} r(XT) &= r(XUV') \\ &\geq r(XUV'V(V'V)^{-1}) \\ &\geq r(XU) \\ &\geq r(XUV') = r(XT) . \end{aligned}$$

Hence, $r(XT) = r(XU)$, where the columns of U are a basis for the column space of T. Any other choice of T must admit the columns of U as a basis. Q.E.D.

*$\mathcal{N}(\cdot)$ denotes the null space.

Theorem I (Miliken (1971)).

The linear contrast $L'\beta$ is estimable, if and only if $r\left[X(I-(L')^{+}L'\right] = r(X) - r(L)$. (4.5.4)

This theorem is equivalent to the theorem of Roy and Roy (1961) since $I - (L')^{+} L' \in \mathcal{N}(L')$.

Theorem II (Miliken (1971)).

The linear contrast $L'\beta$ is estimable if and only if

$$tr\left[X(I-(L')^{+}L')\{X(I-(L')^{+}L')\}^{+}\right] = q-t. \qquad (4.5.5)$$

Proof: Note that the matrix in (4.5.4) is idempotent, hence

$$\begin{aligned} tr\left[X(I-(L')^{+}L')\{X(I-(L')^{+}L')\}^{+}\right] &= r\left[X(I-(L')^{+}L')\{X(I-(L')^{+}L')\}^{+}\right] \\ &= r\left[X(I-(L')^{+}L')\right] \\ &= r(X) - r(L) \\ &= q - t . \end{aligned}$$

Q.E.D.

Theorem (Searle (1965)).

The linear contrast $L'\beta$ is estimable if and only if

$$L' X^{-} X = L' \qquad (4.5.6)$$

for some and hence for every $X^{-} = g_1(X)$.

Proof; If $L' = L'X^{-}X$ then choosing $A' = L'X^{-}$, we have $L' = A'X$, thus $L'\beta$ is estimable.

Conversely, $L' X^{-} X = A' X X^{-} X$, and putting $L' = A'X$

$$\begin{aligned} &= A'X \\ &= L'. \end{aligned}$$

Q.E.D.

Theorem: Let M be a symmetric matrix such that

$\mathcal{C}(X'MV) \subset \mathcal{C}(X'MX)$, in which case

$(y-X\beta)'M(y-X\beta)$, as a function of β , has stationary values.

Let $\hat{\beta}$ be a stationary point.

If $\ell'\hat{\beta}$ is the BLUE of $\ell'\beta$ for every $\ell \in \mathcal{C}(X')$ then it is necessary that

$$r(X'MX) = r(X) \tag{4.5.10}$$

and M is of the form

$$(V + XUX)^{-} + K \tag{4.5.11}$$

for any symmetric choice of g- inverse, where U and K are symmetric matrices such that

$$\mathcal{C}(V;X) = \mathcal{C}(V + XUX'), \tag{4.5.12}$$

$$VKX = 0, \quad X'KX = 0. \tag{4.5.13}$$

Conversely if M is of the form (4.5.11) with 4.5.12) and (4.5.13), then $r(X'MX) = r(X)$ and $\ell'\hat{\beta}$ is the BLUE of $\ell'\beta$ for every $\ell \in \mathcal{C}(X')$.

Proof: Following Rao (1972), we have, on equating the derivative of $(y - X\beta)'(y - X\beta)$ to zero

$$X'MX\,\beta = X'My \ . \tag{4.5.14}$$

This system is consistent since $\mathcal{C}(X'MV) \subset \mathcal{C}(X'MX)$

and $y \in \mathcal{C}(V;X)$, w.p. 1, hence

$$\hat{\beta} = (X'MX)^{-}X'My \tag{4.5.15}$$

is a stationary point. Let $\ell = X'L$. Then $\ell'\hat{\beta}$ is the BLUE of $\ell'\beta$, hence by Searle's theorem

$$L'X(X'MX)^{-}X'MX = L'X.$$

Since L is arbitrary, we have

$$X(X'MX)^{-}X'MX = X, \quad \text{hence} \tag{4.5.16}$$

$$r(X) = r\left[X(X'MX)^{-}X'MX\right] \leq r(X'MX) \leq r(X),$$

i.e., $r(X'MX) = r(X)$, which proves (4.5.10).

If $\ell'\hat{\beta}$ is the BLUE of $\ell'\beta$ for every $\ell \in \mathcal{C}(X')$, then applying the lemma on pp. 317 of Rao (1973a), we have

$$L'X(X'MX)^{-}X'MVZ = 0 \tag{4.5.17}$$

for any L, where Z is a matrix of maximum rank such that $Z \in \mathcal{N}(X')$. Therefore (4.5.17) implies that

$$X(X'MX)^{-}X'MVZ = 0$$

if and only if $X'MVZ = 0$, this implies that

$$VMX = XP \tag{4.5.18}$$

for some P. Now, there exists a symmetric matrix U, such that

$$X'M(V + XUX')MX = X'MX\ . \tag{4.5.19}$$

One choice of U is $(X'MX)^{-}(X'MX - X'MVMX)(X'MX)^{-}$, since

$X'MXUX'MX$

$= X'MX(X'MX)^{-}X'MX(X'MX)^{-}X'MX - X'MX(X'MX)^{-}X'MVMX(X'MX)^{-}X'MX$

$= X'MX - X'MX(X'MX)^{-}X'MXP(X'MX)^{-}X'MX$

$= X'MX - X'MXP(X'MX)^{-}X'MX$

$= X'MX - X'MVMX(X'MX)^{-}X'MX$

$= X'MX - X'MVMX$ by (4.5.16).

Multiplying both sides of (4.5.19) by $X(X'MX)^{-}$ and using (4.5.8) and (4.5.18), we obtain

$$(V + XUX')MX = X. \tag{4.5.20}$$

If $p'(V + X\,U\,X') = 0$, then from (4.5.20) $p'X = 0$, and hence $p'V = 0$ and vice versa, proving (4.5.12). Choosing a $g_{12}(V + X\,U\,X')$ and a symmetric matrix K, let

$$M = (V + XUX')^{-} + K, \tag{4.5.21}$$

substituting (4.5.21) in (4.5.20), we obtain (4.5.13). The converse is easily verified by using lemma 2.

Let $G = X'X + \Sigma = X'X + \Gamma\Gamma'$ where

$G^{-} = g_1(G)$

$S = X'G^{-}\Gamma$

$H = H_o + H_1$

where $H_0 = (I - SS^{+})X'G^{-}$, $H_1 = (S^{+})'\Gamma'G^{-}$.

Alalouf (1975) states and proves a lemma which gives the properties of the matrices G, S, H_0 and H_1. We shall use these results in the sequel and refer the reader to Alalouf (1975) for the proofs. He then uses these results to obtain the following decomposition of the vector space into four virtually disjoint subspaces in terms of four projectors;

Projector	Subspace	Dimension
XH_1	$\mathcal{C}(X) \cap \mathcal{C}(\Gamma)$	$\dim\left[\mathcal{C}(X) \cap \mathcal{C}(\Gamma)\right]$
XH_0	$T_1 \subset \mathcal{C}(X)$ such that $T_1 \quad (\Gamma) = 0$	$r(X) - \dim\left[\mathcal{C}(X) \cap \mathcal{C}(\Gamma)\right]$
$GG^- - XH$	$T_2 \subset \mathcal{C}(X)$ such that $T_2 \cap \mathcal{C}(\Gamma) = 0$	$r(\Gamma) - \dim\left[\mathcal{C}(X) \cap \mathcal{C}(\Gamma)\right]$
$I - GG^-$	$T_3 \subset R^n$ such that $T_3 \cap \mathcal{C}(X,\Gamma) = 0$	$n - r(X_1, \Gamma)$

He then gives the following theorem which justifies the above decomposition.

Theorem. Let G, H, H_o, H_1 be as defined above.

Then

(i) $H_0 Y = H_0 X\beta$ w.p.1. (4.5.22)

(ii) $\mathrm{Cov}\left[XH_1 y, (GG^- - XH)y\right] = 0, \quad \mathrm{Cov}\left[XHy, (GG^- - XH)y\right] = 0$.

(iii) $(I - GG^-)\,y = 0$ with probability one (w.p.1).

Using the decomposition and the previous theorem we obtain the following decomposition of y;

$$y = XH_0 y + XH_1 y + (GG^- - XH)y + (I - GG^-)y \qquad (4.5.23)$$

and note that the fourth term is identically zero and so provides us with no information, while the first and second terms provide information about β. The third term has mean zero and so gives no information about β.

The following lemma gives an expression for the permissible values of β.

Lemma. The vector β satisfies

$$\beta = H_0 y + (I - H_0 X)\,\alpha, \tag{4.5.24}$$

where α is an arbitrary vector and $H_0 y$ is fixed w.p.1

Proof. Using parts (iv) and(vi) of Lemma 5.1 (Alalouf) we have that (4.5.24) is the general solution of (4.5.22).

Theorem. $A'y$ is an unbiased estimator of $L'\beta$ if and only if

$$A'X\left[I - H_0 X\right] = L'\left[(I - H_0 X)\right], \tag{4.5.25}$$

and

$$A'XH_0 y = L'H_0 y\ . \tag{4.5.26}$$

Proof. If $A'y$ is unbiased for $L'\beta$, then for all β satisfying (4.5.24), we have, $A'X\beta = L'\beta$.

Thus $A'XH_0 y + A'X\left[I - H_0 X\right]\alpha = L'H_0 y + L'\left[I - H_0 X\right]\alpha$

must hold for every α. Hence (4.5.25) and (4.5.26). Conversely, if A' and L' satisfy (4.5.25) and(4.5.26), then

$$\begin{aligned} E\left[A'y\right] &= A'X\beta = A'XH_0 y - A'X\left[I - H_0 X\right]\alpha \\ &= L'H_0 y - L'(I - H_0 X)\alpha \\ &= L'\beta. \end{aligned}$$

Q.E.D.

BIBLIOGRAPHY

Aitchison, J., and Silvey, S. D. (1958), "Maximum Likelihood Estimation of Parameters Subject to Restrictions", Ann Math. Statist. 29, 813-828.

Alalouf, I. S (1975), "Estimability and Testability in Linear Models", Ph.D. Thesis, McGill University.

Anderson, T. W. (1958), "An Introduction to Multivariate Statistical Analysis", New York: John Wiley & Sons.

Anderson, T. W. (1961), "Least Squares and Best Unbiased Estimates", Ann. Math. Statist. 33, 266-273.

Bahadur, R. R. (1958), "Examples of Inconsistent Maximum Likelihood Estimates", Sankhyā, 20, 207-210.

Bar-Shalom, Y. (1970), "On the Asymptotic Properties of the Maximum Likelihood Estimate Obtained from Dependent Observations", J.R. Statist. Soc. B., 33, 72-77.

Barnett, V. D. (1966), "Evaluation of the Maximum Likelihood Estimator where the Likelihood Equation has Multiple Roots", Biometrika 53, 1 and 2, 151-165.

Bartlett, D. P. (1915), "The Method of Least Squares", Rumford Press.

Berkson, J. (1956), "Estimation by Least Squares and by Maximum Likelihood"., Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability, Berkeley and Los Angeles, University of California Press, 1956, pp. 1-11.

Bhat, B. R. (1974), "On the Method of Maximum Likelihood for Dependent observations", J.Roy. Statist. Soc. B, 36, 48-53.

Box, G. E. P., and Draper, N. R. (1975), "Robustness to nonnormality of regression tests", Biometrika 49, 1 and 2, 93-106.

Chambers, J. M. (1975), "Fitting non-linear models: numerical techniques", Biometrika 60, 1, 1-10.

Comstock, G. C. (1889), "Method of Least Squares", Ginn and Company.

Conte, S. D., and deBoor, C. (1972),"Elements of numerical analysis: An algorithmic approach", 2nd ed., McGraw-Hill.

Cox, D. R., and Hinkley, D. V. (1966), "A note on the efficiency of Least Squares Estimates", Ann. Math Statist. 37, No. 2, 284-289.

Cramér, H. (1946), "Mathematical Statistics", Princeton University Press.

Daniels, H. E. (1961), "The asymptotic efficiency of a Maximum Likelihood Estimator", Proceedings of the Fourth Berkely Symposium on Mathematical Statistics and Probability, Berkeley and Los Angeles, University of California Press, Vol. 1, p. 151-163.

Draper, N., and Smith, H. (1966), "Applied Regression Analysis", New York: Wiley & Sons.

Edwards, A. W. F. (1972), "Likelihood", Cambridge University Press.

Eicker, F. (1962), "Asymptotic Normality and Consistency of the Least Squares Estimators for families of Linear Regressions", Ann. Math. Statist. 34, 447-456.

Fisher, R. A. (1950), "Contributions to Mathematical Statistics", New York: Wiley & Sons.

Golub, G. H., and Styan, G. P. H. (1973), "Numerical Computations for Univariate Linear Models". J. Statist. Comput. Simul. Vol. 2 253-274.

Graybill, F. A. (1961), "An Introduction to Statistical Models, Vol. 1, McGraw-Hill.

Graybill, F. A., and Marsaglia, G. (1957), "Idempotent matrices and Quadratic Forms in the General Linear Hypothesis", Ann. Math. Statist. 28, 678-686.

Hoerl, A. E., and Kennard, R. W. (1970a), "Ridge Regression: Biased Estimation for Non-Orthogonal Problems", Technometrics, Vol. 12, No. 1, 55-68.

Hoerl, A. E., and Kennard, R. W. (1970b),"Ridge Regression: Applications to Nonorthogonal Problems" Technometrics, Vol. 12, No. 2 69-82.

Hoerl, A. E., Kennard, R. W., and Baldwin, K. F. (1974), "Ridge Regression: Some Simulations", Communications in Statistics, 4(2), 105-123.

Hoadley, B. (1971), "Asymptotic Properties of Maximum Likelihood for the Independent not Identically distributed Case". Ann. Math. Statist. 42, 6, 1977-199.

Hogg, R. V., and Craig, A. T. (1961), "Some Results on Unbiased Estimation", Sankhyā A, 24, 333-338.

Kale, B. K. (1961), "On the Solution of the Likelihood Equation by Iteration Processes", Biometrika 48, 452-456.

Kale, B. K. (1962), "On the Solution of Likelihood Equations by Iteration Processes: The Multiparameter Case", Biometrika 49, 3 and 4, 479-486.

Kendall, M. G., and Stuart, A. (1967), "The Advanced Theory of Statistics", Vol. 2, Hafner.

Lewis, T. O., and Odell, P. L. (1966), "A Generalization of the Gauss-Markov Theorem", J.A. S. A. 66, 1063-1066.

Marsaglia, G., and Styan, G. P. H. (1974), "Equalities and Inequalities for Ranks of Matrices", Linear and Multilinear Algebra, Vol. 2

Milliken, G. A. (1971), "New Criteria for Estimability for Linear Models", Ann. Math. Statist. 42, 5, 1588-1594.

Norton, H. W. (1956), "One Likelihood Adjustment may be Inadequate", Biometrics 12, 79-81.

Prasad, M. S., and Rao, B. L. S. P. (1976), "Maximum Likelihood Estimation for dependent Random Variables", J.I.S.A. Vol. 14, 79-97.

Rao, C. R. (1945), "Generalizations of Markov's Theorem and Tests of Linear hypotheses", Sankhyā 7, 1, 9-19.

Rao, C. R. (1946), "On the linear combination of observations and the general theory of Least Squares", Sankhyā 7, 3, 237-256.

Rao, C. R. (1950), "A theorem on Least Squares", Sankhyā 11, 1, 9-12.

Rao, C. R. (1958), "Maximum Likelihood Estimation for the multinomial distribution with infinite numbers in cells", Sankhyā 20, 3 and 4, 211-218.

Rao, C. R. (1972), "Unified theory of Least Squares", Communications in Statistics 1(1), 1-8.

Rao, C. R. (1973a), "Linear Statistical Inference" 3rd edition. New York: Wiley and Sons.

Rao, C. R. (1973b), "Representations of best linear unbiased estimators in the Gauss-Markov model with a singular dispersion matrix", J. Multivariate Analysis, 3, 276-292.

Rao, C. R. (1973c), "Projectors, Generalized Inverses and BLUE'S", J. Roy. Statist. Soc. B, 36, 442-448.

Rao, C. R., and Mitra, S. K. (1971), "Generalized inverses of matrices and its applications", New York: Wiley & Sons,

Searle, S. R. (1964), "Additional results concerning estimable functions and generalized inverse matrices", J. Roy. Statist. Soc. B, 27, 486-490.

Seber, G. A. F. (1977), "Linear Regression Analysis", New York, Wiley & Sons.

Silvey, S. D. (1961), "A note on Maximum Likelihood in the case of dependent random variables", J. Roy. Statist. Soc. B, 23, 444-452.

Wald, A. (1949), "Note on the consistency of the Maximum Likelihood Estimate", Ann. Math. Statist. 20, 595-601.

Walsh, G. N. W. (1975), "Methods of Optimization", London: Wiley & Sons.

Wolfowitz, J. (1948), "On Wald's proof of the consistency of the Maximum Likelihood", Ann. Math. Statist. 20, 601-602.

Zyskind, G., and Martin, F. B. (1969), "On best linear estimation and a general Gauss-Markov theorem in linear models with arbitrary non-negative covariance structure", SIAM. J. Appl. Math. Vol. 17, 6, 1190-1202.