NONLINEAR STATISTICAL ESTIMATION WITH NUMERICAL MAXIMUM LIKELIHOOD

Gerald Gerard Brown

California University

Prepared for:

Office of Naval Research

October 1974

WESTERN MANAGEMENT SCIENCE INSTITUTE
University of California, Los Angeles

Working Paper No. 222

NONLINEAR STATISTICAL ESTIMATION WITH NUMERICAL MAXIMUM LIKELIHOOD

by

GERALD GERARD BROWN

October, 1974

This work was supported in part by ONR Contract N00014-69-A-0200-4042.

UNIVERSITY OF CALIFORNIA

Los Angeles

Nonlinear Statistical Estimation with Numerical Maximum Likelihood

A dissertation submitted in partial satisfaction of the requirements for the degree Doctor of Philosophy in Management

by

Gerald Gerard Brown

1974

//

CONTENTS

LIST OF FIGURES

VITA

[redacted]--Born, [redacted]

1968--B.A., California State University, Fullerton

1969--M.B.A., Quantitative Methods,
California State University, Fullerton

1969-1973--Lecturer in Quantitative Methods,
California State University, Fullerton

1973-1974--Assistant Professor,
Operations Research and Administrative Sciences,
Naval Postgraduate School, Monterey

PUBLICATIONS

(1) Brown, G. G. and Rutemiller, H. C., "A Sequential Stopping Rule for Fixed-Sample Acceptance Tests," Operations Research, 19, 1971, p.970.

(2) Brown, G. G. and Rutemiller, H. C., "A Cost Analysis of Sampling Inspection Under Military Standard 105D," Naval Research Logistics Quarterly, 20, 1973, p.181.

(3) Brown, G. G. and Rutemiller, H. C., "Some Probability Problems Concerning the Game of Bingo," Mathematics Teacher, 66, 1973, p.403.

(4) Brown, G. G. and Rutemiller, H. C., "Evaluation of Pr{X≥Y} When Both X and Y are from Three-Parameter Weibull Distributions," Institute of Electronic and Electrical Engineers. Transactions on Reliability, R-22, 1973, p.78.

(5) Brown, G. G. and Rutemiller, H. C., "The Efficiencies of Maximum Likelihood and Minimum Variance Unbiased Estimators of Fraction Defective in the Normal Case," Technometrics, 15, 1973, p.849.

(6) Brown, G. G. and Rutemiller, H. C., "Tables for Determining Expected Cost per Unit Under MIL-STD-105D Single Sampling Schemes," American Institute of Industrial Engineers. Transactions, 6, 1974, p.135.

ABSTRACT OF THE DISSERTATION

Nonlinear Statistical Estimation
with
Numerical Maximum Likelihood

by

Gerald Gerard Brown
Doctor of Philosophy in Management
University of California, Los Angeles, 1974
Professor Glenn W. Graves, Chairman

The topics of maximum likelihood estimation and nonlinear programming are developed thoroughly with emphasis on the numerical details of obtaining estimates from highly nonlinear models.

Parametric estimation is discussed with the three parameter Weibull family of densities serving as an example. A general nonlinear programming method is discussed for both first and second order representations of the maximum likelihood estimation, as well as a hybrid of both approaches. A new class of constrained parametric estimators is introduced with numerical methods for their determination.

Structural estimation with maximum likelihood is examined, and a Bernoulli regression technique is presented.

CHAPTER I

STATISTICAL ESTIMATION

A. INTRODUCTION TO THE DISSERTATION

This dissertation is concerned with a class of problems of basic importance in applied statistics - the estimation of parameters in a complicated model where simple closed form estimators do not exist and it is necessary to resort to numerical methods. Many existing numerical approaches prove to be of little practical value in the context of these actual cases because of convergence problems. The main purpose is to develop new numerical techniques by combining recent developments in the theory and practice of optimization with statistical theory and to demonstrate the efficacy of these methods by application to the special class of complicated, highly nonlinear problems arising in statistical estimation. The applications are addressed primarily to maximum likelihood estimation, and the new methods are compared where possible to previous results. The general numerical technique developed is also used to solve a new class of estimation problems with nonlinear constraints on the parameters. The numerical approach is further utilized to provide an alternative to least squares regression, especially for problems with discrete dependent variables.

The present chapter reviews the mathematical foundation for statistical estimation for both density functions and structural models, and provides justification for use of maximum likelihood estimation. Chapter II presents a history of nonlinear programming with both search and ascent methods, with emphasis on numerical performance for highly nonlinear objective functions. Chapter III introduces the

maximum likelihood estimation problem for the parametric Weibull family of density functions. The new techniques of the dissertation are developed and demonstrated. A new class of constrained maximum likelihood estimators is proposed with sample problems. Chapter IV addresses a class of regression models in which the dependent variable is a Bernoulli observation, develops a statistical theory for solutions of the model and gives a numerical example.

B. INTRODUCTION TO STATISTICAL ESTIMATION THEORY

A classical area of intense interest in statistics is the art of using sampling information to make valid inferences about unknown parameters in the distribution of a population under study; this body of technique, motivated by the mathematical theory of statistical estimation put forth by Fisher[86], can be applied in several ways to any given sample producing various estimates of the parameters, and leaving us with the problem of selecting a "good" estimate from among the possibly infinite number of competitors.

An investigator is apt to feel that a "good" estimate is obviously that which is closest to the true parameters. However since the estimator is a mathematical function of the sample (a statistic) it is itself random from sample to sample, so that the attractiveness of a particular randomly distributed estimator will depend upon the long run characteristics described by its sampling distribution. For instance, if the sampling distribution of an estimator for a parameter vector has a great deal of its probability concentrated in a very small neighborhood of the true parameter, and a competing statistic does not, we would probably find the former estimator to be "better" than the latter for purposes of valid inference. That is, the probability of an estimate being close to the true parameter is higher in the former case, so we use that particular method with our sampling information. Unfortunately, there is seldom a guarantee that a statistic will be "good" for every sample, or even that it will produce useful or intuitively acceptable estimates. Therefore, one must choose an estimator on the basis of its long run properties relative to those of feasible alternative estimators and in the context of each application.

In order to formalize some of these concepts of "goodness," let us define the j^{th} observation of an m-dimensional vector, X_j, as

$$X_j = \{ x_{j1}, \ldots, x_{jm} \} , \quad j=1,2, \ldots,n ;$$

with X_j row j of $\underline{X}$, the observation data matrix.

It should be made clear at the outset that if the successive observations in $\underline{X}$ are not random, then we must know the precise nature of the sampling procedure which leads to this non-randomness for the observations, or very little inference is possible. For this reason, $\underline{X}$ is assumed here to result from random sampling from a population with a single set of parameters, T.

For purposes of parametric estimation, we must know, or have assumed hypothetically, the precise mathematical form of the distribution of each observation of the parent population. Therefore, let

$$f_j(X_j,T) ,$$

represent this density, with

$$T = (t_1, \ldots, t_k) ,$$

a set of k columns of unknown parameters to be estimated and f_j non-negative over the region of admissible ranges of X_j and T.

Point estimation, then, is the interpretation of a statistic, $\hat{T}$, computed from $\underline{X}$ as a vector of constants which can be assumed as the inferred value of T; interval estimation is the specification of an interval such that a known proportion of such intervals contain the parameter T.

For simplicity of exposition, let us assume that $f_j = f$ for all j, and momentarily that k = 1. Then, let $\hat{t}(n)$ be a statistic to be used as an estimator of t based on a random sample of size n. It is reasonable to assume that the cost of obtaining the sample is some monotonic increasing function of n, and thus that the economic justification of $\hat{t}(n)$ depends upon how "good" it is as a function of n. In this context some of the following measures of desirability of estimators are proposed as functions of sample size, and thus cost.

1. Existence

It is always necessary to be able to demonstrate that a particular statistic exists with its attendant properties for a given sample space, probability distribution, and so forth.

2. Simple Consistency

A statistic is simply consistent if for any arbitrarily small positive constants c and d there is a sample size N such that

$$Pr[|\hat{t}(n) - t| < c] > 1 - d \ , \ n>N \ .$$

3. Squared Error Consistency

A statistic is said to have squared error consistency if for any arbitrarily small constants c and d and some positive integer n,

$$Pr[(\hat{t}(n) - t)^2 < c] > 1 - d \ , \ n>N \ .$$

Some probabilists view these consistency properties as special cases of stochastic convergence under particular norms. Both types of consistency are desirable in the sense of early discussion in this chapter, producing with high probability values of $\hat{t}(n)$ in a small neighborhood of t, but consistency is achieved at possibly high cost.

4. Bias

The bias, b(n), of the statistic $\hat{t}(n)$ is defined

$$b(n) = E[\hat{t}(n) - t] \ ,$$

with E the expectation operator. If b(n) = 0 for all n,

$$E[\hat{t}(n)] = t$$

and $\hat{t}(n)$ is said to be unbiased.

If b(n) approaches zero as n increases, then $\hat{t}(n)$ is said to be asymptotically unbiased.

Unbiasedness is an intuitively desirable point property, but should not be confused with neighborhood

properties such as consistency; neither property inplies the other. Further, b(n) can sometimes be determined, or estimated, and removed from $\hat{t}(n)$.

5. Variance

The variance of a statistic t(n) is defined

$$V[\hat{t}(n)] = E[\hat{t}(n)^2] - E^2[\hat{t}(n)] = E[(\hat{t}(n) - t - b(n))^2] .$$

This may, or may not, be analytically available depending upon the mathematical form of $\hat{t}(n)$ and f, but it is a characteristic of the sampling distribution of $\hat{t}(n)$ and thus describes long range behavior of $\hat{t}(n)$.

6. Mean Squared Error

The mean squared error of $\hat{t}(n)$ is defined as

$$M.S.E. = E[(\hat{t}(n) - t)^2] = V[\hat{t}(n)] + b(n)^2 .$$

We see that the M.S.E. and variance are identical for unbiased statistics, and that for biased statistics, the M.S.E. exceeds the variance.

7. Likelihood

For independent observations the likelihood of $\hat{t}(n)$ is defined by Fisher[86] as

$$L(\underline{X},t) = f(X_1,t)\bullet\bullet\bullet f(X_n,t) ,$$

and is regarded as proportional to the probability of the occurrence of the vector, $\underline{X}$, given parameter t.

8. Sufficiency

A statistic, $\hat{t}(n)$, is said to be sufficient if it can be shown that the conditional probability distribution, h, of any other statistic, $\tilde{t}(n)$, given $\hat{t}(n)$ does not depend upon the parameter t :

$$h(\tilde{t}(n) \mid \hat{t}(n)) \text{ not a function of } t .$$

Sufficiency for $\hat{t}(n)$ implies that all the sample information concerning t has been exhausted by $\hat{t}(n)$.

Such statistics exist for a very important family of density functions including the exponential, binomial, chi-square, gamma, and normal distributions. As we shall see, a straightforward algorithm may be used to identify sufficient statistics.

9. Completeness

Let $s(X_j)$ be a continuous function of X_j. If $E[s(X_j)]=0$ for every admissible t implies that $s(X_j)=0$ for all X then $f(X_j,t)$ is a complete family of density functions.

10. Minimum Mean Squared Error

It has been shown by Rao[196] and Cramer[58] that under assumptions of regularity the lower bound for M.S.E. of any statistic is

$$M.S.E. = -(1 + db/dt)^2 / E[\partial^2 \ln(L)/\partial t^2] .$$

The regularity assumption disallows discontinuities in f that depend upon t. This bound may or may not be achievable.

For an unbiased statistic, this lower bound for variance is

$$M.V. = -1/E[\partial^2 \ln(L)/\partial t^2] .$$

11. Squared Error Efficiency

A statistic, $\hat{t}(n)$, is relatively efficient if its M.S.E. is less than that of a competitor, $\tilde{t}(n)$, for a given sample size:

$$E[(\hat{t}(n) - t)^2] < E[(\tilde{t}(n) - t)^2] .$$

We can also treat this as an asymptotic property of an estimator. If the inequality ultimately holds for any competitor we simply say that $\hat{t}(n)$ is asymptotically efficient.

This is a very appealing relative measure of the "goodness" of a statistic. It seems reasonable to assume that the cost associated with an error in

estimation is an increasing nonlinear function of the size of the error. For example, the effect of a small error might well be unimportant. A large error, on the other hand, might lead to significant costs due to incorrect decisions based on the estimate. The precise cost-error relationship would be most difficult to specify mathematically. Assuming that the cost is a quadratic function of estimation error gives a cost function that is tractable mathematically, and weights larger estimation errors more heavily than small errors. Thus, with this assumption, a choice of estimators on the basis of relative efficiency becomes a choice of minimum expected cost.

12. Uniqueness

For purposes of inference, it is desirable but often impossible to demonstrate that the statistic used uniquely satisfies its own definition.

13. Asymptotic Normality

An estimator is asymptotically normal if its sampling distribution approaches normality with increasing sample size. This property gives a statistical foundation for making the probability assertions required for interval estimation; it obviates the need, case-by-case, to treat a statistic as a mathematical transformation applied to the random variables in each sample and attempt to use statistical transformation methods to derive a sampling distribution for $\hat{t}(\underline{X},n)$ in closed form. In fact, such an analytic derivation is frequently mathematically impossible.

To use the property of asymptotic normality for interval estimation, we require knowledge of the first two moments of the estimator so that the parameters of the normal distribution may be obtained[4]. In some instances these cannot be obtained analytically, as is shown by Mann, Schafer, and Singpurwalla[161,p.263].

14. Best Asymptotic Normality

A statistic, $\hat{t}(n)$, is Best Asymptotically Normal, B.A.N., if it is simply consistent and $\hat{t}(n) - t$ approaches a normal distribution with zero mean and a variance less than that of any competitor with asymptotic normality over the same open interval for t. (In his introduction of B.A.N. estimators, Neyman[172] gives a more general set of existence conditions in the context of continuous data grouped into classes.) Note that B.A.N. estimators are not necessarily efficient, or unique, but that they are asymptotically unbiased, and of course offer the advantages of asymptotic normality previously discussed.

Finally, with suitable notation adjustments, all these characteristics of point estimators generalize to the multidimensional estimation case, $k > 1$. For instance, the variance should be notationally replaced with a variance-covariance matrix, $\underline{\mathbf{V}}$.

$$\underline{\mathbf{V}} = E[\,(\hat{\mathbf{T}}(n)-\mathbf{T})\;\;(\hat{\mathbf{T}}(n)-\mathbf{T})'\,]$$

C. CHOICE OF ESTIMATOR...DENSITY FUNCTIONS

The art in statistical estimation is as much the choice of an estimator as its mathematical derivation for a given problem. Although a myriad of estimation techniques have been proposed in the literature, only those generally applicable to the problems to be considered here are introduced. Noted by their absence are Bayes estimators, formulated from his idea[15] of using prior information, but which do not apply to a constant vector, **T**, and exist only for very restricted choices of prior multivariate density for **T**, and Minimum Chi-Square estimators, M.C.S., discussed at length by Rao[196], which apply to continuous data grouped into classes, and are very similar in both determination and asymptotic properties to the maximum likelihood estimators, which are presented shortly.

Moment Estimators, $\dot{\mathbf{T}}(n)$, proposed by Pearson[133], are formed by equating the sample moments of $\underline{\mathbf{X}}$ with its theoretical moments stated in terms of the parameters, **T**. The solution for $\dot{\mathbf{T}}(n)$ may not be possible in closed form for many density functions, and $\dot{\mathbf{T}}(n)$ is not necessarily unique for any given sample, however Pearson introduced a large family of special distributions which yield solutions for $\dot{\mathbf{T}}(n)$. Moment estimators are usually consistent in both the simple and squared error sense, asymptotically normal, but not b.A.N., and can be efficient only when the variance of $\underline{\mathbf{X}}$ dominates higher moments of f, as is true with the normal distribution. In general, $\dot{\mathbf{T}}(n)$ has few advantages over common competitors for any particular density function, f,

and the disturbing habit of frequently producing outrageously bad estimates, even inadmissible ones. The use of moment estimators by Pearson and others has been largely restricted to the more specialized problem of choosing both a mathematical form for f when none is known, and estimation of resulting parameters.

Sufficient statistics, $\check{T}(n)$, have been demonstrated by Fisher[86] and Neyman[174] to exist for any density for which the likelihood function may be partitioned into

$$L(\underline{X},T) = H(\check{T}(n),T)\ K(\underline{X})$$

with H an exclusive nontrivial function of $\check{T}(n)$ and T, and K free of terms or constraints involving T. A condition implying existence of a sufficient statistic, $\check{T}(n)$, is that f belong to the Koopman-Pitman exponential family[142,186] such that f may be stated

$$f(X_j,T) = \exp[\,p(T)\,m(X_j) + s(X_j) + q(T)\,]$$

with p(T) a nontrivial continuous function of T, $s(X_j)$ and $m(X_j)$ continuous functions of X_j, $dm/dX_j \neq 0$, and the range of X_j independent of T.

Sufficient statistics are of strong intuitive appeal since they demonstrably use all of the sample information available. The algorithm for finding a sufficient statistic is straightforward, leading immediately either to the establishment of such a statistic, or a proof that no sufficient statistic exists [128,p.231, 141,p.26]. Unfortunately, sufficient statistics are not necessarily consistent, unbiased, or efficient.

Any nontrivial one-to-one transformation of $\check{T}(n)$ is also sufficient for T. Therefore, whenever possible we choose an estimator from this infinite family of sufficient statistics in order to achieve one or more additional desirable properties such as consistency, minimum variance, or most often unbiasedness.

A Minimum Variance Unbiased Estimator, M.V.U.E., discussed by Rao[195] and Blackwell[24], is always a function of the sufficient statistic, and is found as the conditional expectation of any statistic which is unbiased for T, given the sufficient statistic, $\check{T}(n)$. The M.V.U.E., when it can be derived via the conditional density required, is necessarily simply and M.S.E. consistent, and the most efficient unbiased statistic for any sample size. Further, if the density function is complete, the M.V.U.E. is unique[128,p.229]. The mathematical details of deriving the M.V.U.E. are arduous, but the statistic is desirable especially for small samples where bias and/or M.S.E. are high for most competitors. A minimum M.S.E. statistic provides a tradeoff by minimizing the sum of variance and squared bias, and can be preferable to the M.V.U.E. when unbiassedness is not absolutely essential. Unfortunately, minimum M.S.E. statistics are only rarely derivable for finite sample sizes, and when found often correspond to the M.V.U.E. result. For instance, the sample mean from a normal distribution can be shown to be both a M.V.U.E. and minimum M.S.E. statistic.

Maximum Likelihood Estimators, M.L.E., suggested by Fisher[86], are found by maximizing the likelihood function $L(\underline{X},T)$ by choice of T. These intuitively appealing estimators, $\hat{T}(n)$, can often be derived in closed form by differential calculus, and always exist under mild regularity conditions. Although $\hat{T}(n)$ is frequently biased for small samples, it is asymptotically unbiased, B.A.N.,

and simply and squared error consistent as shown by Wald[225,224]. It is also asymptotically efficient, ultimately achieves the minimum variance bound, and can be shown to be a function of the sufficient statistic, if one exists. Even for relatively small samples, the M.L.E. can be more efficient than the M.V.U.E., as has been shown by Brown and Rutemiller[31].

M.L.E. also have an important invariance property. For any non-trivial function of T, u(T), with a single-valued inverse,

$$\hat{u}(T) = u(\hat{T}(n)) \ .$$

For example, invariance permits transformations to reduce bias without sacrifice of other desirable M.L.E. properties.

This property is an indispensable tool in mathematical modelling. Since parametric estimation is usually performed only as a preliminary part of a larger investigation, invariance is crucially important, permitting M.L. point estimates to be unconditionally introduced into any admissible function of the associated parameters, with the function directly inheriting all the desirable M.L. properties. This permits analysis of complex hierarchical systems to be conducted in a straightforward manner.

Asymptotic normality for all M.L.E. makes them very useful for interval estimation, especially in the multivariate sense. Unfortunately, M.L.E. can not, in general, be guaranteed to be unique, although uniqueness can be established on a case-by-case basis. Although numerical determination of M.L.E. can at worst be exceedingly difficult in practice, the "good" properties of these estimators make them so singularly attractive in the general field of statistical estimation as to motivate the

investigation in this thesis.

D. CHOICE OF ESTIMATOR...STRUCTURAL MODELS

Suppose we examine a model in which the population mean is not strictly a function of T, but rather a particular mathematical function of the population parameters, T, and some observed constants, $\underline{X}$. If we define our sampling process to be the measurement, with some random error, of observations, Y, from populations whose parameters depend on $\underline{X}$ and T, then a problem which results is the estimation of the parameters, T, based on the sample

$$\{Y , \underline{X}\} ,$$

by use of the relationship

$$Y = Y(T, \underline{X}, e) ,$$

and known information about the nature of the error, e. This technique is known as regression.

One example of such a model is classical linear regression, where

$$Y = \underline{X}T + e .$$

Since $Y-\underline{X}\tilde{T}(n)$ is the sample estimation error in the model for the estimator $\tilde{T}(n)$, the usual approach to this estimation is to assert a quadratic cost function and minimize the scalar sum of squared deviations

$$(Y-\underline{X}T)' (Y-\underline{X}T)$$

by choice of T. This technique was first suggested for use in interpolation of planetary data by Legendre[150]. Provided that $\underline{X}'\underline{X}$ is non-singular, which requires $n > k$,

this quadratic objective function has a unique solution,

$$\underset{\sim}{T}(n) = (\underline{X}'\underline{X})^{-1} \underline{X}'Y .$$

This Least Squares, L.S., estimator is attractive to use for linear models. The L.S. solution is the best linear unbiased estimator, B.L.U.E., in the sense that among all unbiased linear combinations of $\underline{X}$ this estimator has minimum variance regardless of the distribution of e. Gauss[95] has shown that when e is normal the L.S. solution always maximizes the joint density

$$f(y_1|X_1,T) \cdots f(y_n|X_n,T) .$$

This remarkable demonstration both anticipates the later discovery of M.L.E. and shows that in the normal case, the linear model has a single solution which is both L.S. and M.L.E. The distributional theory for interval estimation in the linear model is presented by Cochran[51], and is based on the unique class properties of the multivariate normal density, which is closed for affine transformations, convolutions, and linear mixtures of normals, and the class of chi-square distributions of quadratic normal forms, which is closed for convolutions.

The assumption of normality for e and linearity are crucial to the L.S. approach, since for non-normal, or non-linear models the distributional results fail. In fact, the specification of a quadratic cost criterion for L.S. minimization is not necessarily justifiable in all applications; for instance, mean deviation, or minimax (Tchebycheff) deviation might sometimes be more reasonable.

A general M.L.E. approach to regression focuses attention on the density of e to specify the likelihood

function

$$L(Y|\underline{X},T) ,$$

which is maximized by choice of T. It is not necessary to derive $L(Y|\underline{X},T)$ from $Y(T, \underline{X}, e)$ if one can state the density directly as in the case of Bernoulli regression examined in Chapter IV. The M.L.E. solution has all the properties under conditions mentioned previously, regardless of the form of the model, although those that are asymptotic are achieved more slowly for highly non-linear models or extraordinary distributions for e. Sprott and Kalbfleisch[217] have examined for some specific models the robustness of the assumption of asymptotic normality made for several finite sample sizes.

E. SUMMARY: JUSTIFICATION OF M.L.E.

As we have seen, the M.L.E. usually have, for large samples, all the desirable properties of an estimator. They almost always exist under very mild regularity conditions, asymptotically they are consistent, unbiased, efficient, B.A.N., achieve the Cramer-Rao minimum variance bound, and they are sufficient statistics whenever such statistics exist. They can often be derived in closed form by differential calculus, and in other cases, the estimator may be solved for by numerical techniques.

When point estimates of functions of parameters are required in a mathematical model, it is pointless to choose estimators for their "good" properties unless the function will also possess those properties. In practice, the M.L.E. are the only available estimators with so many desirable properties that are all invariant under such transformations. As mentioned earlier, this invariance property of M.L.E. is vital in complicated problems where parametric estimation is only the first step of investigation.

Best of all, M.L.E. provide a distributional basis for interval estimation which does not depend upon simplifying assumptions such as those required for the L.S. approach. This is fortunate, since in models which are non-normal, non-linear, or, more often, both, the M.L.E. provide the only reasonable estimation alternative. Also, for the classic linear normal model, the M.L.E. provides the L.S. solution.

For small samples, the M.L.E. have many of the good estimator properties, and are often the best statistic available. Their M.S.E. is frequently the best among

competitors, even for very small sample sizes. The M.L.E. are extremely useful in small sample estimation as a starting point for seeking better statistical estimators for particular density functions. The M.L.E. are always derived by exactly the same method, requiring less intuition, skill, or plain luck than the intricate schemes leading sometimes to, for instance, an M.V.U.E. In some statistics texts, in fact, M.L.E. are the only estimators introduced since they are generally easy to find and usually produce better estimates than other methods[156,p.162].

Among alternative estimators for any given problem, the M.L.E. nearly always provide a very good property set that gets better very quickly with increasing sample size, and becomes asymptotically best. For those cases in which the M.L.E. must be determined numerically, a potentially difficult nonlinear programming problem results.

CHAPTER II

NUMERICAL TECHNIQUES OF ESTIMATION

A. INTRODUCTION TO NONLINEAR NUMERICAL ESTIMATION

In the previous chapter we have proposed a statistically desirable nonlinear estimation method, M.L.E., which leaves us with the problem

$$\underset{T}{MAX}\{\ L(T)\ \}\ .$$

The form of L depends upon the model used. Estimation of density function parameters for f gives

$$L(\underline{X},T) = f(X_1,T)\bullet\bullet\bullet f(X_n,T)\ ,$$

and estimation of parameters for a structural model gives

$$L(Y|\underline{X},T) = f(Y|X_1,T)\bullet\bullet\bullet f(Y|X_n,T)\ .$$

In either case, L is known to be a highly nonlinear function of the decision variables, T. Since $\underline{X}$ and Y are treated as constants in these two models, they will not be included in the further notation of this chapter, so that both estimation models may be treated at once. Thus

$$L(T) = f_1(T)\bullet\bullet\bullet f_n(T)\ .$$

Mathematical constraints may be present for the parameters. These may be simple numerical range

constraints, upper and lower bounds, for instance

$$l_i \leq t_i \leq u_i \ , \ i=1, \ldots, k \ ,$$

or more complicated joint functions of T, equality constraints of the form

$$g_1(T) = 0 \ ,$$

or inequality constraints such as

$$g_2(T) \leq 0 \ .$$

The set of both types of constraints is referred to collectively as

$$g'(T) = \{g'_1(T), g'_2(T)\} \ ,$$

$$g(T) \leq 0 \ .$$

We refer to the conditioned set of all values of T which simultaneously satisfy the constraint set, $g(T)$, as the feasible region for T, and values of T within that region are called feasible points. A particular constraint that is exactly satisfied by T (a row of $g(T)$ exactly equal to zero) is said to be active. If, for all possible pairs of two feasible points, T_1 and T_2, the convex combination

$$T = aT_1 + (1 - a)T_2 \ , \ 0 \leq a \leq 1 \ ,$$

is also feasible, then the feasible region is called convex.

For M.L.E. problems, there are frequently simple

numerical bounds placed on **T**. These are usually included to insure the definition of a valid density function, f. However, general mathematical constraints are seldom present. For this reason we will initially emphasize the unconstrained M.L.E. problem and the techniques available for its solution.

The first step in formulating an M.L.E. problem for solution is usually the replacement of the likelihood function, L, by its logarithm, ln[L]. It is easy to see that

$$\underset{T}{MAX}\{\ L(T)\ \}\ ,\ \text{and}\ \underset{T}{MAX}\{\ \ln[\,L(T)\,]\ \}\ ,$$

are both achieved by the same value of **T**, since the logarithm is a monotonic increasing function of its argument. The log-likelihood function becomes

$$\ln[\,L(T)\,] = \ln[\,f_1(T)\,]+\ldots+\ln[\,f_n(T)\,]\ ,$$

This reformulation usually gives an alias for L(T) which is a mathematically simpler function. For instance, members of the Koopman-Pitman family of density functions are remarkably easier to deal with in this form. This is advantageous for both analytic and numerical work. For instance, since L(T) is the product of n sample likelihoods, its value for many problems, especially for large n, can numerically violate the expressible range of floating point representation on a particular digital computer.

We henceforth treat L(T) as the objective function, in either the likelihood, or aliased log-likelihood form. Further, we assume where necessary that L(T), and thus f(T), are continuous, twice differentiable functions of **T** at interior points. This is a very weak restricting assumption

for M.L.E. models, which very seldom have discrete parameters, T, and rarely have non-differentiable density functions (poles, etc.) for realistic problems in which M.L. estimation is attempted. It is not necessary in a mathematical programming sense to emphasize the statistical relationship of the M.L.E. and sample size, so it is assumed notationally that

$$\hat{T} = \hat{T}(n) \ .$$

A stationary point of $L(T)$ is characterized[21] by the necessary condition that the gradient vanish at $\hat{T}$,

$$\nabla L(\hat{T}) = \partial L(T)/\partial T \,|_{T=\hat{T}} = 0 \ .$$

Necessary conditions for a local maximum are that

$$\nabla L(\hat{T}) = 0 \ ,$$

and that the symmetric Hessian matrix,

$$\underline{H} = \{h_{ij}\} = \{\partial^2 L(T)/\partial t_i \partial t_j\} \ ,$$

be negative semidefinite at a stationary point, $\hat{T}$; that is, for any vector z not identically zero,

$$z' \underline{H}(\hat{T}) z \leq 0 \ .$$

A vanishing gradient and negative definite Hessian

$$\nabla L(\hat{T}) = 0 \quad \& \quad z' \underline{H}(\hat{T}) z < 0 \ ,$$

provide sufficient conditions for a local maximum of L.

If the Hessian can be shown to be negative definite for all feasible points T, then L is said to be concave[19], and a stationary point, $\hat{T}$, is the unique global maximum. Other characterizations of stationary points of L are possible; these other cases are of little general use and usually require further assumptions for identification of maxima, such as higher-order derivatives[208].

Characterizations of extrema of $L(T)$ in the presence of equality constraints requires that the gradient vanish while all the equality constraints simultaneously hold. Lagrange[147] expressed these conditions by introducing an r-dimensional vector of arbitrary multipliers, u_1, and augmenting the objective function of the problem to include the constraints, giving

$$\underset{T,u}{MAX}\{L(T) - u_1' g_1(T)\} ,$$

which, as previously shown, is stationary if

$$\nabla_{T,u}[L(\hat{T}) - u_1' g_1(\hat{T})]=0 \ \& \ r \leq k ,$$

and a local maxima under conditions for the Hessian similar to those for the unconstrained problem, but modified by the dimensionality adjustment. John[135], and later Kuhn and Tucker[146], have generalized the necessary conditions to inequality constraints as follows, letting u_2 be a vector of multipliers associated with $g_2(T)$:

$$\nabla_T[L(\hat{T}) - u_2' g_2(\hat{T})]=0 ,$$

with

$$g_2(\hat{T}) \leq 0, \quad u_2 \leq 0, \quad u_2' g_2(\hat{T}) = 0 \ .$$

The last condition is referred to as complementary slackness.

For maximization problems subject to mixed constraints, with multipliers defined

$$u' = \{u_1', u_2'\} \ ,$$

necessary conditions for a local constrained maximum are:

$$\nabla_T [L(\hat{T}) - u' g(\hat{T})] = 0 \ ,$$

$$g_1(\hat{T}) = 0, \quad g_2(\hat{T}) \leq 0, \quad u_2 \leq 0, \quad u_2' g_2(\hat{T}) = 0 \ .$$

Local sufficiency for these conditions further requires that the constrained objective function be locally concave, that all nonlinear inequality constraints be convex, and that all equality constraints be linear. It may be possible to generalize local sufficient conditions, subject to the Kuhn-Tucker restrictions, for nonlinear equality constraints.

John[135] actually developed conditions requiring that the objective function also have a multiplier, and Kuhn and Tucker[146] qualified admissible constraint sets to those without singularities on the boundary such as an outward pointing cusp, or other nonlinear degeneracy; in these cases, the multiplier proposed by John is positive, and can in fact be normalized to unity. Their development defines the Lagrangian objective function

$$\mathcal{L}[T,u] = L(T) + u'g(T) ,$$

and specifies that if a stationary point $(\hat{T},u^*)$ is also a saddle pcint, that is

$$\underset{T}{MAX}\, \mathcal{L}[T,u^*] = \mathcal{L}[\hat{T},u^*] = \underset{u}{MIN}\, \mathcal{L}[\hat{T},u] ,$$

that under the mild assumptions, the point $(\hat{T},u^*)$ is a solution to both the primal and dual problems, given respectively at the left and right above. This also suggests that methods for solution of the primal problem can sometimes profit from information gained by simply examining, cr shifting emphasis completely to the dual. We might interpret the primal optimization process as maximization subject to feasibility with respect to constraints and the dual optimization process as minimization of infeasibility, subject to a stationary primal profit criterion.

Further characterizations under varying sets of assumptions and useful simplifying qualifications have been given by Mangasarian and Fromovitz[159], Arrow and Enthoven[6], Arrow, Hurowicz and Uzawa[7], Kortanek and Evans[143], and Wilde[230,231].

For many likelihood functions, $\hat{T}$ may be determined in closed form as a stationary point of L by differential calculus. In such cases, demonstration of extremality and uniqueness proceed directly by analytic means as previously discussed.

In general, however, the stationary points of L must be derived iteratively by the numerical methods of nonlinear

programming. The general M.L. estimation problem has rather distinctive features in this respect. The number of decision variables, or parameters, is usually very small, seldom more that three for density functions and ten for structural models. The objective function and especially its gradient are highly nonlinear, expensive to evaluate numerically, and difficult to compute precisely. These problems are exacerbated by large sample sizes. The constraints are usually of relatively simple form, often just numerical bounds on T.

B. METHODS OF NUMERICAL OPTIMIZATION

The nonlinear programming methods which may be used for M.L. estimation are all iterative schemes with the following features. An initial value of T, T_0, must be specified or guessed by the investigator. An iteration mechanism then chooses a step-size and direction for determining the sequence

$$T_0, T_1, \ldots, T_m ,$$

such that

$$L(T_i) > L(T_{i-1}), \quad i=1,2, \ldots, m.$$

Finally, a set of termination states is specified. Termination criteria commonly include a maximum value of m. A stalling criterion can be used for tolerance of resolution, with d a vector of arbitrary small constants,

$$|T_m - T_{m-1}| < d .$$

A performance criterion can be employed to insure acceptable distinguishability, or marginal improvement,

$$L(T_m) - L(T_{m-1}) > \text{minimal gain.}$$

The ideal iteration scheme is a totally automatic

algorithm in that the global solution is reached in a finite number of steps without necessitating human intervention. Unfortunately, no single method realistically qualifies on this basis, especially if we define finiteness in terms of exhausting a reasonable computer budget. Also, a global solution does not always exist in the strict sense for all M.L. problems. In practice, even the attainment of a local maximum can be delightful.

A good iteration algorithm should not require excessive computation time for termination. Neither should it demand brilliant intuition, or extraordinary good fortune, on the part of the user. Problem specificity of good iteration performance is also undesirable, unless for demonstrable cause of an apparent nature general enough to advise prior choice of the method.

The taxonomy of iteration schemes identifies direct search methods as those which achieve gains by experiment with evaluation of the objective function, L(T). Ascent methods, on the other hand, require local derivative information to calculate a priori where each following evaluation of the objective function should take place. Ascent methods may be further subclassified as either direct ascent, which seek immediate gains at each iteration, or indirect ascent, which seek at each step to achieve the necessary conditions for a maximum. Note that ascent methods include those using finite difference approximations to derivatives. Distinguishing between these two classifications is at times most difficult, since the systematic experimental achievement of increases in the objective function, L(T), by varying the argument, T, with a direct search scheme is highly suggestive of cognizance of differential information indicative of an ascent method. This interminable classification problem is obviated by the plausible defense of nomistic innocence. Several classical

techniques of both types that are available for finding t(n) when k=1, for instance golden section search, regula falsi, and so forth[232,193], are not discussed here.

C. DIRECT SEARCH METHODS

The Hooke-Jeeves pattern search method[129], perhaps the simplest technique known, is a maximization scheme based on direct evaluation of L(T). Given a starting point, T_0, and stepsize, s_0, the iteration sequence proceeds by varying each element of T by one step in each direction and evaluating the objective function, keeping each respective element of T at the value which lead to a maximum. Thus, T_1 will be at a corner of the k-dimensional hyperrectangle defined by $T_0 \pm S_0$. The scheme proceeds similarly until no further gain seems possible, in which case the stepsize is halved, the process repeated to no gain, stepsize halved again, and so forth until termination is recognized within a small enough neighborhood.

Several heuristic modifications have been proposed, including a ridge-following "memory" for acceleration of stepsize when an element of s continues step-to-step to exhibit no change in sign while sequential gains are made[129], a sequential transformation of coordinates in order to minimize parameter interaction and separate the effects of steps on the approximately orthonormalized problem, linear minimizations along estimated conjugate directions, a restart procedure for avoiding local minima and stalling, parallel tangent acceleration suggested by Shah, Buehler, and Kempthorne[210], quadratic approximation with an interpolating polynomial over the local search lattice, and introduction of random numbers to avoid dead ends for the search. Such ad hoc modifications are found in Fletcher[87], Zangwill[239], Rosenbrock[206], Powell[190],

and Davies[66], who also describes response surface direct optimization schemes encountered in experimental design.

The simplex method, introduced by Spendley, Hext, and Himsworth[216], generalized by Nelder and Mead[171], and generally referred to as the simplical scheme so as not to confuse it with the linear programming algorithm, uses k+1 points defined as a simplex in the k-dimensional search space. At each iteration a new point is created to replace the point associated with the minimum value on the simplex by reflection of the minimum point via a ray through the centroid of the other points over a distance determined by a reflection constant. A possible dimensional collapse of the simplex is avoided by special logic, and acceleration and convergence are achieved, respectively, by expansion of the maximum point on the simplex on a ray from its centroid, or contraction of the minimum point on the simplex on a ray toward the centroid.

This ingenious technique works much like the pattern search methods examined above, and will almost always terminate eventually by converging to a local maxima. Modifications of the scheme are possible with random perturbations to mitigate near linear dependencies in the simplex and to avoid final convergence to a local maximum. Numerical bounds can be accomodated on the parameters. Box[27] found the simplical scheme superior to pattern search and Rosenbrock's[206] method, and introduced the "complex" search method, which is a generalization of the simplical scheme to admit a convex inequality constraint set. Richardson and Kuester[199] have published another constrained simplical program. One weakness of the method is the requirement for an interior T_0, but Noh[177], has further generalized the complex search for equality constraints and non-interior starting points. Box reported

that for his simple models, objective function evaluations commonly required 1000 times as long as the complicated step selection logic. Parkinson and Hutchinson[181] discuss the relative merits of variations of the simplical approach.

Although simplical schemes seem to work in practice, even for difficult problems, no acceptable formal proof of convergence has yet appeared. The theoretical difficulty seems to lie in (unconstrained) counter examples which can be constructed and for which the method should not terminate. For instance, see the cases given by Shere[211] for the program presented by Richardson and Kuester[199]. Realistically, however, confrontation of such special cases is highly unlikely. On the other hand, it is true that dimensional collapse is a continuing theoretical and numerical hazard in the presence of constraints. Finally, it should be noted that these are scarcely substantive criticisms of the method when it is used for adaptive process control, as it was originally intended.

Direct search methods which attempt to reliably achieve global maxima have been proposed by Brooks[29], Bocharov and Fel'dbaum[25] and Page[180]. These treat the objective function as an unknown but deterministic response to the argument, T. The optimization proceeds by sequentially partitioning mutually exclusive and exhaustive regions for interior T over which the first two moments of the objective function are estimated to discriminating precision by random sampling or numerical quadrature over a k-dimensional lattice, and a hypothesis test is performed to select the better region, which is in turn bisected on the next step. The iteration ceases when an acceptably small region is selected.

It is important to note the difference between these area evaluation methods and simple random point sampling.

Without the partitioning scheme and sequential area estimation and hypothesis tests, these methods degenerate to the infamous Las Vegas technique.

Each area selection method suffers from a non-parametric probability of excluding the region containing the global optimum at some intermediate decision step and thus of unreliably reporting a surrogate, nonlocal suboptimal solution. Geometric features such as an isolated peak with steep slopes and a shallow base can evade detection and can be caused by a poor choice of initial feasible region for interior points.

Several authors, notably Clough[50], Cooper[55], Hartley[119], Hartman[120], Liau, Hartley, and Sielken[154], and Zakharov[238] have developed statistical strategies for region sampling and evaluation and conducted experiments with standard objective functions. They report limited success in actual applications. None of the applications include a problem typical of M.L.E.

High frequency oscillations and other irregularities which thwart other search techniques are smoothed and thus mollified by this area approach. This smoothing characteristic and the academically appealing global strategy suggest the technique for finding a reasonable starting domain for interior points for some other search mechanism, especially if the latter iteration converges only in a close neighborhood of a maximum, or if the objective function is pathological. Some experimentation has shown, however, that excessive objective function evaluations were necessitated for relatively small, uncomplicated sample problems.

D. ASCENT METHODS

Most indirect schemes are characterized by an iteration of the form

$$T_i = T_{i-1} + a\underline{M}^{-1}s \ , \ i=1,2, \ldots, m \ ,$$

with a positive scalar step length, a, an iteration matrix $\underline{M}^{-1}$, and a vector of directional gradient information, s. For instance, the first-order method of steepest ascent first described by Cauchy[45], and later by Courant[56], Curry[59], and Levenberg[153], uses

$$\underline{M} = \underline{I} \ , \ s = \nabla L(T) \ ,$$

and chooses the stepsize a as a suitable positive constant to increase L(T) along the ray

$$T_{i-1} + a\underline{I}\nabla L(T) \ .$$

a may be chosen to produce a maximum along the ray by direct evaluation, regula falsi, quadratic approximation, or simply to produce any gain. This method ultimately terminates at a local maxima, but often converges with slow performance, especially along curved rising ridges for which it hem-stitches with agonizing progress.

Further discussion of ascent methods is given by Goldstein[101] and Ramsay[194]. Powell[190] and Brent[28] give first-order ascent schemes using difference approximations for derivatives, with due attention to the

numerical and theoretical consequences of such substitution.

A second-order scheme, the Newton-Raphson method, applies

$$\underline{M} = -\underline{H}(T), \; s = \nabla L(T), \; a = 1 \; ,$$

for which convergence termination depends upon negative definiteness of $\underline{H}(T)$. This condition on $\underline{H}(T)$ is usually guaranteed only over a small neighborhood satisfying the Lipschitz condition discussed by Henrici[123], which in essence requires that $L(T)$ behave nearly linearly in the vicinity of a maxima. The rate of convergence for problems that do successfully terminate is quadratic above the noise level of machine calculations and it follows rising ridges well. However, this second-order scheme is renowned for its propensities to seek saddle points and follow ridges out of the vicinity of the feasible region. Also, computing $\underline{H}$ can be prohibitively expensive and imprecise for $L(T)$, requiring, as it commonly does, k^2 very extensive n-sums of complicated nonlinear transcendental terms. (Not to speak of the debugging effort in checking program logic and algebra.) Goldstein and Price[103], have suggested approximation of $\underline{H}$ by finite differences on $L(T)$ in these cases. Error analysis of the Newton-Raphson scheme is given by Lancaster[148].

Many methods have been proposed to give convergence rates like those of Newton-Raphson and dependability of steepest ascent. Usually these involve forming an iteration matrix, $\underline{M}$, by various means in the interests of assuring positive definiteness over the largest neighborhood.

The conjugate gradient method, invented by Hestenes and Stiefel[126], applies an ingenious one-step memory by

modifying the steepest ascent iteration to the recursion

$$s_i = \nabla L(T_i) + (\|\nabla L(T_i)\|/\|\nabla L(T_{i-1})\|)\, s_{i-1} ,$$

with

$$s_0 = \nabla L(T_0)/\|\nabla L(T_0)\| .$$

This scheme avoids the notorious hem-stitch stalling of the steepest ascent method, even permitting finite convergence proofs for quadratic objective functions. The ortnogonalized gradient vectors, and the conjugacy and linear independence of the steps is achieved at very little cost, without requiring maintenance of second order information, such as H(T). Thus, second order convergence can often be achieved at very little additional computational cost. The method was suggested for solving linear systems by Hestenes and Stiefel[126] and implemented for nonlinear objective functions later by Fletcher and Reeves[90]. A complete development is given by Hestenes[124,125]. A convergence discussion and modifications to the method are given by Daniel[60].

Fisher[86] gives the second-order method of scoring, also discussed by Rao[197], which is specific to problems in which a log-likelihood function is maximized, and is identical to the Newton-Raphson method, except that the Hessian is replaced by its expectation,

$$\underline{M} = E[-\underline{H}(T)] ,$$

where $\underline{M}$ is called the information matrix, which Kendall and Stuart[141,p.56] show to always be positive definite. We

see that the final iteration matrix for this scheme, $\underline{M}(\hat{T})^{-1}$, is the Cramer-Rao bound for regular M.L.E. Vandaele and Chowdhury[223] give some computational examples and suggest some minor modifications for this approach. This method requires a formal derivation of the expectation of some very complicated transcendental sums in the Hessian matrix. An example will serve to illustrate the scope of this problem later.

Both the theoretical and numerical performance of these iteration methods can be improved by appropriate affine transformation of the problem. For instance, see the recent investigation of Amor[3]. Other techniques can be applied to insure positive definiteness for $\underline{M}$. Various spectral decompositions of $\underline{M}$ may be used. Determination of eigenvectors and associated eigenvalues of the real symmetric matrix $\underline{M}$ is possible by several methods reviewed by Schwarz, Rutishauser and Stiefel[209], along with square root and Cholesky decompositions. Although diagonalization and orthonormalization of $\underline{H}$ will eliminate local parameter interaction, the neighborhood over which the result holds is quite small for non-quadratic problems, making the transformation of questionable value when performed at the high expense of the eigen-analysis. If the condition number of $\underline{H}$ is defined as the ratio of the absolute values of the largest to the smallest eigenvalues, then a measure results of both topological distortion from an idealized k-dimensional response sphere about T, and the difficulty with which $\underline{M}$ will be accurately inverted[127,78,133].

Advocates of the transformational approach have even proposed introducing constraints on the eigenvalues of $\underline{M}$, for instance, replacing negative eigenvalues by their absolute values, and near-zero values by a small constant was proposed by Greenstadt[108] for maximization with a

Newton-Raphson-like scheme. With some difficulty we can momentarily visualize the presence of a large condition number implying the existence of a long ridge or trough oriented with the eigenvector associated with the eigenvalue in the denominator. This is a good situation for a second-order iteration scheme if the ridge is convex, which is the case when the eigenvalue in the denominator of the condition number is positive. This eigenvalue constraint method, and other similar proposals, attempt to mask the concave ridges and saddle points which are also attractive in the second-order iteration. Booth and Peterson[26] discuss such geometric inference at length.

A reasonable compromise is the simple scaling of $\underline{M}$, analgous to the creation of a correlation matrix from a covariance matrix. Let a scaling of $\underline{M}$ be performed by

$$\underline{M}_s = \{m_{ij}/|m_{ii}m_{jj}|^{1/2}\} ,$$

with singularities $m_{jj}=0$ replaced in the computation by 1. This can ease the burden of computing spectral decompositions for the iteration matrix, and it can reduce internal loss of numerical precision in the iteration scheme.

In the same vein, a normalized gradient is sometimes applied

$$\nabla L(T) = \nabla L(T)/||\nabla L(T)|| ,$$

to keep computations numerically stable and place the scaling burden on the scalar stepsize, a. Even though these transformation methods are always available and sometimes useful, they are not emphasized in this presentation for

simplicity. This is appropriate in part since the investigator should always take care to reasonably scale any problem regardless of the method employed to solve it.

Levenburg[153] proposes a scheme which has since been generalized and machine implemented by Marquardt[164]. In the development, a method is sought which will behave like steepest ascent in regions not local to the solution, and like Newton-Raphson when the solution is approached. The iteration matrix is chosen

$$\underline{M} = -\underline{H}(T) + m\underline{I} \ ,$$

with m a positive constant. We see that no matter how ill-conditioned $\underline{H}$ is, a suitably large choice for m will give a numerically nonsingular iteration matrix.

(The nonsingularity of $\underline{M}$ is more apparent if we momentarily consider the convex combination

$$\underline{M} = -(1-a)\underline{H}(T) + a\underline{I}, \ 0 \leq a \leq 1 \ . \ )$$

For m=0, this Marquardt-Levenburg heuristic is the Newton-Raphson method, and for m large this approaches the steepest ascent method. Marquardt gives a heuristic for modifying m by a multiplicative expansion/reduction factor on the basis of algorithm performance. A more formal method of determining m was later put forth by Smith and Shanno[212], along with facility for handling linear constraints by the projected gradient method of Rosen[203].

Marquardt also introduces a useful termination criterion for tolerance of resolution. With "|...|" denoting a k-vector of absolute values, this is

$$|T_m - T_{m-1}| \leq 10^{-5}(T_{m-1} + 10^{-3}) .$$

This might be restated

$$|T_m - T_{m-1}| \leq 10^{-d}(T_{m-1} + 10^{-n}) , \quad (d + n)\ \ln_2 10 < b,$$

with d the number of significant digits of desired resolution. b is the number of bits in the floating point mantissa of the computer used, modified by the noise level for one or two's complement arithmetic.

Another school of thought attempts to achieve second-order convergence without evaluating $\underline{H}$ at each step of the iteration. The iteration matrix, $\underline{H}$, is assiduously, and hopefully, maintained as a negative definite substitute for $\underline{H}^{-1}$. Such variable metric methods, introduced by Davidon[63], and discussed by Broyden[35], are in reality more computationally efficient indirect ways of approximating the Hessian matrix by differencing as suggested earlier by Golstein and Price[103]. These approaches work by adding a correction matrix at each step

$$\underline{M}_i^{-1} = \underline{M}_{i-1}^{-1} + \underline{C} ,$$

with $\underline{C}$ derived in several ways. Define

$$\Delta T = T_i - T_{i-1} = a\underline{M}^{-1}s = a\underline{M}^{-1}\nabla L(T) ,$$

$$\Delta(\nabla L(T)) = \nabla L(T_i) - \nabla L(T_{i-1}) ,$$

$$\Delta s = \Delta(\nabla L(T)) \ ,$$

and

$$d = \Delta T - \underline{M}_{i-1}^{-1} \Delta s \ ,$$

then a rank-one correction for the iteration matrix, $\underline{M}^{-1}$, is

$$\underline{C} = dd'/\Delta T'd \ ;$$

there are others, for instance see Householder[130,p.123].

A rank-two correction for the iteration matrix, developed by Davidon, and Fletcher and Powell[89], gives

$$\underline{C} = \Delta T \Delta T'/\Delta T'\Delta s - \underline{M}_{i-1}^{-1} \Delta s \Delta s' \underline{M}_{i-1}^{-1} / \Delta s' \underline{M}_{i-1}^{-1} \Delta s \ .$$

An inverse rank-one correction proposed by Powell[191] and Bard[12] uses

$$c = \Delta s - \underline{M}_{i-1}^{-1} \Delta T \ ,$$

to give

$$\underline{C} = cc'/\Delta T'c \ .$$

Powell[191] suggests using

$$\underline{M}_i^{-1} = \underline{M}_{i-1}^{-1} + \underline{C} \ ,$$

while Bard suggests

$$\underline{M}_i^{-1} = \underline{C} .$$

These rank-one methods have also been discussed by Greenstadt[109], Fiacco and McCormick[83,p.170], Cantrell[43], Miele and Cantrell[168], Cragg and Levy[57], Forsythe[92], Myers[170], and many others, largely with the objective of finding a stepsize with minimal expenditure and avoiding singularities in $\underline{M}$. Lill[155] presents a computer program with some of these features. Rank-two and other variable metric schemes have been examined by Bard[11], Davidon[64], Goldfarb[99], Matthews and Davies[165], Brown and Dennis[33], and Broyden[36,229,230], who gives evidence against using transformations on the problem when in a near neighborhood to the solution under pain of stalling the algorithm. On the other hand, Oren and Luenberger[178,179] propose a self-scaling variable metric class of algorithms with claims of excellent performance.

These methods have been compared with others intended for the more general problem of solving a simultaneous set of nonlinear equations by Barnes[13], Daniel[61], and Broyden[34,39]. For contrast, it is also instructive to review earlier work by Davidenko[62], and Wolfe[235].

A further modification of second-order schemes is introduced in two excellent papers by Stewart[218], and Gill and Murray[97], in which the gradient is estimated by differences, and sequential approximations of the Hessian are made with great care in an attempt to balance truncation errors, loss of numerical precision, and ill-conditioning in the iteration matrix. These authors mention the numerical singularities that can occur in the iteration matrix despite theoretical guarantees to the contrary. Gill and Murray propose the spectral decomposition known as Cholesky

factorization for representing the symmetric Hessian. For $\underline{L}$ a lower triangular matrix and $\underline{D}$ a diagonal matrix, the factorization produces

$$\underline{M} = \underline{L}\underline{D}\underline{L}' \ .$$

Definiteness for $\underline{M}$ is then assured by the careful monitoring of diagonal elements of $\underline{L}$ and $\underline{D}$.

Jones[136] gives a factorization for Marquardt's scheme. Jones, Ross[207] and Bard[12], give comparisons of the various indirect iteration schemes, finding the Marquardt and Davidon-Fletcher-Powell methods better in most test problems. Brooks[30] gives a review of ealier unconstrained methods, as do Dennis[71], Powell[192], Spang[215] and Kowalik and Osborne[144].

E. METHODS WITH CONSTRAINTS

General constraints on the optimization problem have already been defined notationally along with characterizations of optima under these conditions. Algorithms permitting constraints are classifiable by the admissable form of the constraints and the associated objective function. For instance, a linear constraint set can be treated with classical linear programming, L.P., methods if the objective function is approximated linearly. Note that the L.P. includes mechanisms for the determination of interior points, T_i, given any starting value for T_0. Frank and Wolfe[93] present such a first-order algorithm, for linearly approximated objective functions, stated for step i:

$$\underset{T}{MAX} \ \nabla L(T_{i-1})'T_i \ ,$$

which is solved via a standard L.P. step (treating $\nabla L(T_{i-1})$ as a fixed parameter vector), reapproximated, and so forth. Other similar approaches to the problem have been proposed by Wolfe[236] who uses the Kuhn-Tucker conditions to formulate a L.P. for a quadratic objective function, while Beale[16,17] and Zangwill[240] imbed the objective function evaluation within the L.P. mechanism. Non-convex problems have been approached similarly with decomposition techniques discussed by Zangwill[242]. A primal-dual method is given by van de Panne and Whinston[222].

Nonlinear equality constraints may be implicitly combined with the objective function by the use of Lagrange multipliers, as discussed earlier, to produce an

unconstrained equivalent maximization problem. For nonlinear inequality constraints, a set of active, or basic, constraints is kept in the objective function and used via the current multipliers, or their estimates, to give feasible directions for each iteration, either along an active constraint, or toward the interior region. These modifications are discussed for gradient methods and quadratic objective functions by Markowitz[163], Theil and van de Panne[220], and Lemke[152]. General problems with mixed nonlinear constraints are examined by Rosen[203,204], Davies[65], Zoutendijk[244], Forsythe[91], Goldstein[100], and many others. Goldfarb[98] gives a generalization of the Davidon - Fletcher-Powell second order method to accomodate mixed linear constraints. Greenstadt[107] presents a local deflected gradient method.

Nonlinear constraints may also be explicitly added to the objective function by the use of penalty functions, an idea attributed by some to Courant[56], recently suggested by Carroll[44] and generalized by Fiacco and McCormick[82,83,81]. For example:

$$\underset{T}{MAX} \quad L(T)$$

$$s.t.\ g_1(T) \leq 0\ ,\ g_2(T) = 0\ ,$$

is restated with "interior" penalty functions

$$\underset{T}{MAX}\ L(T) + c/g'_1(T)\,1 - g'_2(T)g_2(T)/c^{1/2}\ ,$$

with c a scaling parameter, and 1 a summing vector. As an interior point approaches any constraint, the objective function is distorted. This sequential unconstrained optimization technique, S.U.M.T., solves a sequence of

monotonically less internally distorted problems by decreasing c to a noise level. We see that a formal basis of active constraints need not be maintained, although logic should be included to permit numerical evaluation of the ratios in the objective function as they approach indeterminate limits. Sequential relaxation of the penalties will ultimately terminate with an interior solution, or for problems with active constraints in their final solution, a termination occurs in a close neighborhood of the undistorted solution. Great care must be taken in constructing the S.U.M.T. iteration so as to properly scale, or "tune," the constant, c.

Zangwill[241] gives an "exterior" penalty function formulation

$$\underset{T}{MAX} \quad L(T) - cg'(T)g(T) \ ,$$

with $g(T)$ the subset of constraints from $g_1(T)$ and $g_2(T)$ violated by the current solution, and c a positive constant sequentially increased maximization-to-maximization to an arbitrarily large terminal value. While this method admits any starting solution, T_0, there is an added burden of maintaining a current index set for violated constraints.

Many other variations have been proposed for penalty methods, notably by Camp[42], Butler and Martin[41], Goldstein and Kripke[102], Stong[219], Pomentale[189], Fiacco[79], Fiacco and Jones[80], Kowalik, Osborne and Ryan[145], and Beltrami and McGill[18].

Finally, cutting plane algorithms introduced by Kelley[140] for a linear objective function and nonlinear

constraints, and by Cheney and Goldstein[47] and Wolfe[237] for strictly concave objective function and constraints, and a constraint set which is convex, involve successive introduction of auxilliary variables and constraints to a sequence of linearly bounded problems. Such strategies can lead to cumbersome dimensionality and numerical overhead even for relatively small problems.

The texts by Hadley[111], Fletcher[88] and Mangasarian[158], give extensive development of the various constrained algorithms.

F. SUMMARY: AN EFFICIENT GENERAL TECHNIQUE

Convergence proofs are widely published for most of the numerical optimization methods presented thusfar. For instance, Zangwill[243] develops several representative theorems, each with its set of simplifying assumptions and necessary conditions. However, even for a "nice" problem (convex, quadratic, and so forth) these mathematical demonstrations all implicitly depend at some point upon exact arithmetic, and are thus weakened by finite numerical precision of floating point operations on a digital computer. As an example, the effect of numerical, or random, perturbations on an iteration matrix and thus its inverse is largely a mathematical problem that is not well understood. Perhaps one is better off to adopt a passive view. An undesirable, but nontheless terminal state of an iteration algorithm is always possible due to mathematical and numerical instabilities. This is the motivation of the "terminal state" approach taken here, rather than a "convergence" point of view.

The relative computational success of an algorithm in practice often becomes a more important criterion for its selection than theoretical rate-of-convergence. Further, one must usually trade off the degree of automation of a method (the amount of monitoring and "tinkering" required for each application) with efficiency stated either in terms of solution expense or the probability of termination at a stationary point that is optimal. In short, sufficient proof is performance, and it is never general.

Along these lines, it can be dangerous to attempt to generalize the results of computational experiments on "standard" functions, such as those discussed by Rosen and Suzuki[205], to a complicated application (very nonlinear,

high dimensionality, etc.). One of the reasons for the lack of literature on comparisons of algorithm performance on real problems is the incredibly high cost of conducting such competitions, measured in exhausted computer accounts and man-hours expended in preparation. Other factors contributing to the paucity of published comparisons include the sheer volume of data required to display the results meaningfully, and the proprietary nature of both the problems and implemented algorithms.

There are some refreshing exceptions, such as the report of Friedman and Pinder[94] who state that the complex method performed better for their application than S.U.M.T., deflected gradients, or pattern search.

Graves[104] gives a description of a general nonlinear programming algorithm developed and used for several test cases and applied to a more complicated minimum fuel guidance problem. Graves and Whinston[106] present both analytic evidence, and experimental results for convergence of the method on a set of problems given by Colville[54]. Hatfield and Graves[122] give another favorable application, and Clasen, Graves and Lu[49] describe a set of large munitions mix allocation problems upon which the method is successfully applied. Hense, the efficacy of the Graves algorithm has been established on both theoretical and empirical grounds over a period of ten years' use.

The method is based on the optimization of a sequence of local linear programming problems arising from the first order approximation of the objective function and constraints in the neighborhood of a current solution. Significantly, the program has general provisions for the practical implications of certain numerical characteristics of finite precision digital computers and for vagaries in the behavior of general nonlinear constraint sets.

The efficient performance of the method is probably most directly attributable to the speed, precision and compactness of the linear programming algorithm exercised; the mutual primal-dual representation of linear programs given by Balinski and Gomory[10] and extended and completely implemented by Graves[105] provides both impressive solution speed and a format within which the local linear problem may be easily and effectively manipulated to deal with various inconsistencies and parameterizations which arise during the course of solution of a nonlinear problem.

Linearly constrained problems are handled expeditiously by this general nonlinear method. Modifications of the iteration mechanism are facilitated by program features permitting, for instance, incorporation of a second order representation of the problem given in [106]. Externally supplied routines may be used to monitor the progression of solutions, provide starting tableaus from previous solutions, perform specialized input/output functions, and so forth. As further evidence of the adaptability of the Graves philosophy in nonlinear programming, Hatfield[121] demonstrates an efficient conjugate gradient method for a linearly constrained nonlinear programming problem.

The general method uses the fundamental linear approximation theorem presented and proved by Buck[40, p. 180] for continuous differentiable functions, $g(\mathbf{T})$. If, for some $\mathbf{T}_0$ and single constraint, $g(\mathbf{T})$,

$$g(\mathbf{T}_0 + \Delta\mathbf{T}) = g(\mathbf{T}_0) + \nabla g(\mathbf{T}_0)'\Delta\mathbf{T} + \text{rem}(\mathbf{T}_0, \Delta\mathbf{T}) ,$$

then

$$\lim_{\Delta\mathbf{T}\to 0} \text{rem}(\mathbf{T}_0, \Delta\mathbf{T})/|\Delta\mathbf{T}| = 0 ,$$

is approached uniformly for feasible T_0. The iteration proceeds by solution of the local linear programming problem

$$\underset{\Delta T}{MAX} \quad \nabla L(T_0)'\Delta T ,$$

$$s.t. \quad \nabla \underline{G}(T_0)'\Delta T \leq -g(T_0) - \bar{K}r ,$$

with r a vector of positive constants representing the directional linear approximation error estimated from the most recent iteration, and initialized $r=0$. $\bar{K}$ is a parametric adjustment constant, used to control the rate of solution of the local problems.

Three numerical bounds are imposed on the algorithm. The first is an upper bound on the variables of the dual program

$$\underset{Y}{MIN} \quad Y'[g(T_0) + \bar{K}r] ,$$

$$s.t. \quad Y'\nabla \underline{G}(T_0) = \nabla L(T_0) ,$$

$$Y \geq 0 ;$$

which is stated

$$B_1 \geq Y'1 .$$

This condition insures that the optimal bases of the sequence of local linear programs do not approach singularity arbitrarily closely while remaining nonsingular, and is used in lieu of the Kuhn-Tucker constraint qualifications.

The primal variables are bounded

$$B_2 \geq ||\Delta T||$$

so that numerical range constraints on T may be incorporated algebraically into solutions without inclusion in the constraint set, g(T), and to preclude local numerically unbounded solutions.

The last bound,

$$B_3 \geq MAX\{|REM(T_0,\Delta T)|\} ,$$

gives an upper limit for the linear approximation remainder terms. This error bound is used with B_1 during the progress of the algorithm to control the parametric adjustment for infeasibilities in local linear programs via the constant $\bar{K}$.

A zero level, e, is also provided as a "noise" limit for numerical computations within the program. This is a very important feature in several ways. For instance, the pivotal transformations use e to control accumulation of truncation errors. Most important, constraints are considered to be satisfied when

$$g(T) \leq 0 + e .$$

This is a subtle feature. Some thought about numerical evaluation of nonlinear functions bounding the feasible region reveals that apparent inconsistencies caused by loss of real precision could lead to incorrectly concluding that an infeasibility has been encountered, when in fact T is in a feasible e-neighborhood of, for instance, an equality constraint. Remember, too, that the local linear program

will, when finally applied to maximizing the objective function, seek basic extremal solutions on the boundary of inequality constraints as well. Thus this e-relaxation is a fundamental technique. In the lexicon of Iverson[134], we treat constraint boundaries as being "fuzzy."

A sequence of consistent local linear programming problems is solved by constructively treating violated linear approximations of constraints as objective functions in subproblems. If for some intermediate solution $\bar{K}$ has been parametrically reduced to

$$\bar{K} = e/[B_3(B_1 + 1)] ,$$

and there still remains a violated constraint $g(T_0)$, not a member of active constraints $\overset{*}{g}(T_0)$ with associated dual variables $\overset{*}{Y}$, such that

$$\overset{*}{Y}'\overset{*}{g}(T_0) \geq -g(T) - e ,$$

then an unresolvable inconsistency is reached as a terminal state.

A terminal optimal solution is recognized when a local linear program exhibits a dual solution with

$$Y'g(T_0) \geq -e .$$

A finite convergence proof for this technique requires,

as always, restricting assumptions about the nonlinear functions in the problem. However, a terminal optimal solution to a local linear program is a stationary point for the original objective function, $L(X,T)$. The possibility of termination at a stationary saddle point cannot be ruled out, but experience has shown such a result to be very rare for real problems.

A second order representation of the primal problem can often be expected to converge more quickly in the neighborhood of a stationary point than the first order "gradient" formulation. To achieve the higher order representation, we create an expanded nonlinear program by introducing the first order stationary conditions as constraints. This expanded representation introduces the dual variables explicitly and uses

$$T^{*} = \{T, \overset{\#}{T}\} ,$$

so that the reformulation yields

$$\underset{T^{*}}{MAX} \; [\nabla L(T) - \nabla \underline{G}(T)' \overset{\#}{T}]' T + g(T)' \overset{\#}{T}$$

$$\text{s.t.} \qquad g(T) \leq 0 ,$$

$$\nabla L(T) - \nabla \underline{G}(T)' \overset{\#}{T} \leq 0 ,$$

$$T^{*} \geq 0 .$$

We define $\overset{\#}{\underline{H}}(T)$ as the three dimensional matrix of Hessians for the constraint set, with "column j" the Hessian of $g_j(T)$. Now, with $T^{*} = T^{*}_0$, the parameterized local linear program becomes

$$\underset{\Delta T^*}{MAX}[\nabla L(T_0) + \{\underline{H}(T_0) - \overset{\#}{\underline{H}}(T_0)\overset{\#}{T}_0\}T_0]'\Delta T + [g(T_0) - \nabla\underline{G}(T_0)T_0]'\overset{\#}{\Delta T}$$

$$\text{s.t.} \qquad \nabla\underline{G}(T_0)\Delta T \leq -g(T_0) - \bar{K}r_1 ,$$

$$[\underline{H}(T_0) - \overset{\#}{\underline{H}}(T_0)\overset{\#}{T}_0]\Delta T - \nabla\underline{G}(T_0)'\overset{\#}{\Delta T} \leq -\nabla L(T_0) + \nabla\underline{G}(T_0)'\overset{\#}{T}_0 - \bar{K}r_2$$

In the special case of unconstrained problems, the local linear program becomes

$$\underset{\Delta T}{MAX}\ [\nabla L(T_0) + \underline{H}(T_0)T_0]'\Delta T ,$$

$$\text{s.t. } \underline{H}(T_0)\Delta T \leq -\nabla L(T_0) - \bar{K}r_2 ,$$

and the direction of ascent, neglecting the parameterization term $\bar{K}r_2$, becomes

$$T = -[\underline{H}(T_0)^{-1}]'\nabla L(T_0) ,$$

which is the familiar Newton-Raphson result when the Hessian is of full rank.

It should be noted that the present linear programming approach is more robust than the classical Newton-Raphson process precisely because it will continue to function in the presence of a singular or near singular Hessian. Also, the relaxation features of the method introduced via the parameterization of the right hand side with $\bar{K}r$ generally have a salutary effect on the rate of convergence for a constrained first order or any second order representation

of the problem.

CHAPTER III

ESTIMATION FOR THE WEIBULL DENSITY FUNCTION

A. INTRODUCTION TO THE PARAMETRIC WEIBULL FAMILY

A density function has been proposed for describing breaking strength in materials and later formally introduced by Waloddi Weibull [227,228,229] for use in fitting many types of physical data from various academic and industrial fields of interest. It is his reasonable claim that there is very seldom sound theoretical basis for applying any particular density to real data. He therefore advises choice of a relatively simple density function which seems to fit with empirical observations, and "stick to it as long as none better has been found[229,p.293]."

The Weibull density was originally parameterized

$$T' = \{ a, b, c \} ,$$

and given the form

$$f_1(x,T) = (b/a)(x - c)^{b-1} \exp\{-[(x - c)^b/a]\} ;$$

$$a, b > 0; \; x > c .$$

A reparameterization gives the equivalent

$$f_2(x,T) = (b/a)[(x - c)/a]^{b-1} \exp\{-[(x - c)/a]^b\} ;$$

$$a, b > 0; \; x > c .$$

In this form of the three parameter Weibull, a is known as the "scale parameter," b as the "shape parameter," and c as the "location parameter."

The flexibility of the Weibull family of densities via choice of the shape parameter, b, is illustrated in Figure 1 for arbitrary location parameter, c, and unit scale parameter, a. The chameleonic nature of this family is discussed by Lehman[151]. Its robust adaptability for data fitting have made it a popular candidate in such applications. Indeed, with b=1 the Weibull simplifies to the two parameter exponential family, and when b=2 the Rayleigh family results. Figure 2 shows the Rayleigh family of densities arising from the Weibull with b=2, arbitrary location parameter, c, and several values of the scale parameter, a.

By design, the Weibull density is a perfect algebraic differential, and its reliability function is defined

$$R_2(x,T) = \int_x^{\infty} f_2(x,T)\ dx = \exp\{-[(x - c)/a]^b\} ,$$

and the distribution function follows:

$$F_2(x,T) = \int_c^x f_2(x,T)\ dx = 1 - \exp\{-[(x - c)/a]^b\} .$$

Keen interest in the Weibull family comes from reliability applications and the statistics of extremes. Gumbel[110] gives a derivation of a form of the Weibull family under the name "Type III asymptotic distribution of the smallest extreme." Also, reliability theory leans heavily upon the concept of "hazard rate," which is defined

$$h(x,T) = f_2(x,T) \; / \; R_2(x,T) \; .$$

This is interpreted as the instantaneous failure rate of a functioning electronic device or physical component under service stress.

Many statistical studies are made under hypothetical conditions of decreasing, constant, or increasing hazard rate. The flexible Weibull family can exhibit all three. In fact, another derivation of the Weibull density comes immediately from the assumption of

$$h(x,T) = (b/a)[(x - c)/a]^{b-1} \; ,$$

as the mathematical form for the hazard rate function.

In an excellent introduction to reliability theory, Mann, Schafer and Singpurwalla give many references to applications in the open literature using the Weibull, and state: "Recently, the Weibull distribution has emerged as the most popular parametric family of failure distributions." [161,p.127]

FIGURE 1

WEIBULL DENSITIES FOR a=1, b=1.0(0.5)4.0

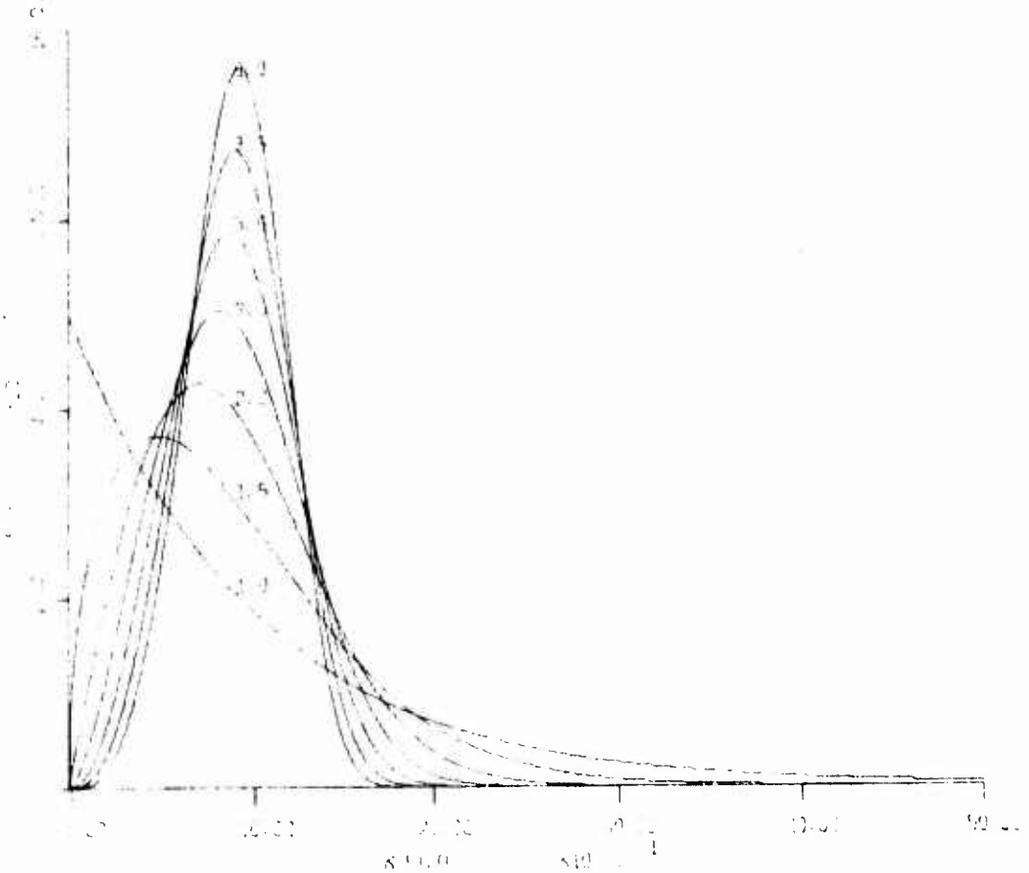
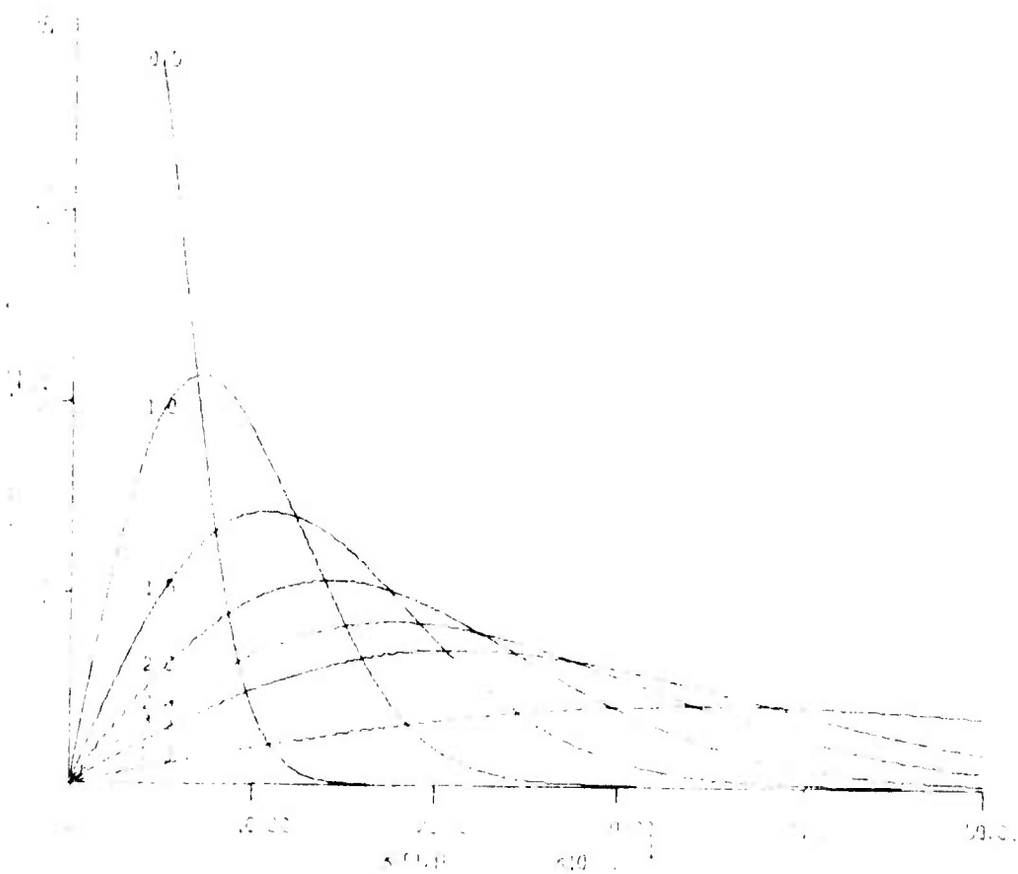


FIGURE 2

WEIBULL DENSITIES WITH a=0.5(0.5)3.0,5.0, b=2



B. ESTIMATION ALTERNATIVES

The most popular statistical techniques for parametric estimation in the Weibull family are graphical estimation, the method of moments, and M.L.E. The first method needs little discussion, however the latter two demand some mathematical development for proper evaluation, especially for the M.L.E. For the present we will restrict our attention to complete random samples, and the most general case of all three parameters unknown. The literature is rife with examples of estimation for subsets of the parameters, although numerical details are often scarce. The most prolific author on the subject is Mann[160;161,p.185ff.] who gives extensive references. Dubey[75,73,72,76] has also made many contributions. Also see the cases given by Menon[166] and Smith and Dubey[213]. Generalizations of special cases for subsets of Weibull parameters have been given by Dubey[77] and others. An excellent discussion of the entire topic is given by Rockette[201].

Graphical estimation, used by Weibull[229] and described by Berrettoni[22] and Kao[139] relies on some prior knowledge of parameter values and the fact that the reliability function, in the form proposed by Weibull,

$$R_1(x,T) = \exp\{-[(x - c)^b/a\} ,$$

can be transformed to

$$\ln\{-\ln[R_1(x,T)]\} = b \ln(x - c) - \ln(a) .$$

A value for the location parameter, c, is asserted with

reference to the first order statistic in a sample. Then the empirical reliability function is plotted on a "log-log" ordinate scale versus a "log" abcissa of the displaced sample values, X - c. If the resulting points fall in a nearly straight line, then b is estimated as its slope and a is found from the intercept, ln a. Otherwise, another value of c is tried, and so forth.

Obviously, this subjective estimation method leaves much to be desired statistically. However, it can be carried out with tools no more formidable than an extensive table of logarithms, and it has served adequately for decades. Of course, a L.S. approach to this transformed problem is also possible, but the results are statistically comparable to the manual method.

The method of moments can be used to estimate Weibull parameters. Although a moment generating function for the Weibull cannot be given in closed form algebraically, the central moments are defined

$$m'_q = E[x^q] = \int_c^\infty x^q f_2(x,T)\ dx\ ,$$

from which we derive for the q^{th} moment the partial sum

$$m'_q = \sum_{i=0}^{q} \binom{q}{i} c^{q-i} a^i \Gamma(1+i/b)\ .$$

The first two moments about the mean are

$$m_1 = c + a\Gamma(1+1/b)\ ,$$

and

$$m_2 = a^2 [\Gamma(1+2/b) - \Gamma^2(1+1/b)] .$$

Obviously, it is impossible to solve these equations explicitly for the parameters, although an iterative solution is possible by elimination of one parameter. Surprisingly, however, the skewness

$$m_3/m_2 =$$

$$\frac{\Gamma(1+3/b) - 3\Gamma(1+2/b)\Gamma(1+1/b) + 2\Gamma^3(1+1/b)}{[\Gamma(1+2/b) - \Gamma^2(1+1/b)]^{3/2}}$$

and kurtosis

$$m_4/m_2^2 =$$

$$\frac{\Gamma(1+4/b) - 4\Gamma(1+3/b)\Gamma(1+1/b) + 6\Gamma(1+2/b)\Gamma^2(1+1/b) - 3\Gamma^4(1+1/b)}{[\Gamma(1+2/b) - \Gamma^2(1+1/b)]^2}$$

are strictly functions of the shape parameter, b. Each can individually yield estimates of b by reference to a simple tabulation [198], depending upon the judgement of the data analyst.

Given $\hat{b}$, one may sequentially obtain moment estimators of the scale parameter

$$\hat{a} = \{m_2/[\Gamma(1+2/\hat{b}) - \Gamma^2(1+1/\hat{b})]\}^{1/2} ,$$

and location parameter

$$\hat{c} = \bar{x} - \hat{a}\,\Gamma(1+1/\hat{b}) \ .$$

Unfortunately, these moment estimators do not have many desirable properties. Dubey[74] has investigated the efficiency of moment estimators for Weibull parameters in restricted cases. The most common complaint is that the estimate of location parameter, $\hat{c}$, regularly violates the simple numerical constraints of the definition of the Weibull random variable. Most often this would be

$$0 < c < x_{[1]} \ .$$

When a violation occurs, it is not clear that any reasonable method exists for adjusting $\hat{c}$ and "backing out" the change through the other moment equations. The most common practice in these cases is to replace $\hat{c}$ by the violated bound, and conveniently disregard the effect upon the moments of the solution.

C. MATHEMATICAL PRELIMINARIES

M.L.E. for the parametric Weibull family may ostensibly be numerically performed with almost any of the unconstrained techniques introduced in Chapter II. Therefore, in order to compare the merits of various approaches, the following development gives the necessary mathematical basis for consideration of any of the search or ascent methods.

The log likelihood function for the Weibull family $f_2(x,T)$ is

$$L(X,T) = \ln\left[(b/a)^n \prod_{i=1}^{n} [(x_i - c)/a]^{b-1} \exp\{-[(x_i - c)/a]^b\} \right] ;$$

$$a,\ b > 0,\ c \geq x_{[1]} ,$$

with $x_{[1]}$ the first order statistic. This gives

$$L(X,T) = n \ln[b/a] + (b - 1) \sum_{i=1}^{n} \ln[(x_i - c)/a]$$

$$- \sum_{i=1}^{n} [(x_i - c)/a]^b .$$

The gradient of the log likelihood, $\nabla_T L(X,T)$, has the three elements

$$\nabla_a = (b/a) \{ \sum_{i=1}^{n} [(x_i - c)/a]^b - n \} ,$$

$$\nabla_b = n/b + \sum_{i=1}^{n} \ln[(x_i - c)/a] - \sum_{i=1}^{n} [(x_i - c)/a]^b \ln[(x_i - c)/a],$$

$$\nabla_c = -(b - 1) \sum_{i=1}^{n} [1/(x_i - c)] + (b/a) \sum_{i=1}^{n} [(x_i - c)/a]^{b-1} .$$

The symmetric Hessian matrix for the family, $f_2(x,T)$, parameterized by

$$T' = \{ a, b, c \} ,$$

is defined as

$$H(T) = \{h_{ij}\} = \partial^2 L(X,T)/\partial t_i \partial t_j ,$$

and is given by

$$h_{11} = (b/a^2) \{n - (b+1) \sum_{i=1}^{n} [(x_i - c)/a]^b\} ,$$

$$h_{12} = (1/a) \{-n + \sum_{i=1}^{n} [(x_i - c)/a]^b + b \sum_{i=1}^{n} [(x_i - c)/a]^b \ln[(x_i - c)/a]\} ,$$

$$h_{13} = -(b/a)^2 \sum_{i=1}^{n} [(x_i - c)/a]^{b-1} ,$$

$$h_{22} = -n/b^2 - \sum_{i=1}^{n} [(x_i - c)/a]^b \ln^2[(x_i - c)/a] ,$$

$$h_{23} = -\sum_{i=1}^{n}(x_i - c)^{-1} + (1/a)\{b\sum_{i=1}^{n}[(x_i - c)/a]^{b-1}\ln[(x_i - c)/a] + \sum_{i=1}^{n}[(x_i - c)/a]^{b-1}\},$$

$$h_{33} = -(b-1)\{\sum_{i=1}^{n}(x_i - c)^{-2} + (b/a^2)\sum_{i=1}^{n}[(x_i - c)/a]^{b-2}.$$

We see immediately that the three parameter Weibull family, $f_2(x,T)$, is not of the Koopman-Pitman form admitting surficient statistics. Therefore, search for an M.V.U.E. is pointless.

Special cases of the Weibull family have already been mentioned for known b=1 (exponential) and b=2 (Rayleigh). The former case is covered exhaustively in the literature. M.L. estimation for the two parameter exponential is nonregular, requiring use of

$$\hat{c} = x_{[1]}.$$

We shall hencefforth rule out values of b<1, since the likelihood function is unbounded as the location parameter, c, approaches $x_{[1]}$ and thus the Weibull density is inappropriate for use in the M.L. estimation.

Rockette[201] analyzes the other cases with b>1, for various subsets of the parameters known. If both a and b, or a and c, are known, solution of the appropriate remaining gradient element gives a unique M.L. estimator, as is

verified by examination of the numerical behavior of the applicable conditioned Hessian term.

If b and c are known, a Jacobian transformation

$$v = (x - c)^{b} ,$$

gives an exponential density for v with well known properties ircluding uniqueness for $\hat{a}$.

If only b is known, the resulting solution for the M.L.E. is unique. McCool[157] shows that if only c is known, the remaining M.L.E. are unique. We shall see later that knowledge of a is of little value, since $\hat{a}$ can be derived as a function of $\hat{b}$ and $\hat{c}$.

Proceeding with the general three parameter case, it will be reassuring for purposes of validation to show that the expectation of the gradient satisfies

$$E[\nabla_{T} L(X,T)] = 0 .$$

This derivation, and others which follow, require definition of several mathematical functions and identities. For scalar z>0, the gamma function is defined as:

$$\Gamma(z) = \int_{0}^{\infty} y^{z-1} e^{-y} dy ,$$

with the useful relation

$$\Gamma(z+1) = z\Gamma(z) ,$$

and the derivatives $\Gamma'(z)$ and $\Gamma''(z)$, are defined by:

$$\Gamma^{(i)}(z) = \int_0^{\infty} \ln^{(i)}(y)\, y^{z-1} e^{-y}\, dy \ , \quad i=1,2,\ldots \ .$$

Also, there are tabulations given by Gauss[96], H. Davis[67] and P. Davis[69] of the digamma (Psi) function

$$\Psi(z) = \nabla_z \ln\Gamma(z) = \Gamma'(z)/\Gamma(z)$$

which has the recursive property

$$\Psi(z+1) = \Psi(z) + 1/z$$

and the trigamma function, with tabulations presented by H. Davis[68] and P. Davis[69], defined

$$\Psi'(z) = \nabla_z[\nabla_z \ln\Gamma(z)] = \Gamma''(z)/\Gamma(z) - [\Gamma'(z)/\Gamma(z)]^2$$

$$= \Gamma''(z)/\Gamma(z) - \Psi^2(z) \ .$$

This will permit algebraic substitution using

$$\Gamma'(z) = \Gamma(z)\Psi(z)$$

and

$$\Gamma''(z) = \Gamma(z)[\Psi'(z) + \Psi^2(z)] \ .$$

Now, the expectations of the gradient, $\nabla_T L(\mathbf{X},\mathbf{T})$, are

$$E[\nabla_a] = (nb/a)\Gamma(2) - nb/a = 0 \ ,$$

$$E[\nabla_b] = n/b + (n/b)\Gamma'(1) - (n/b)\Gamma'(2)$$

$$= (n/b)[1 + \Gamma(1)\Psi(1) - \Gamma(2)\Psi(2)]$$

$$= (n/b)[1 + \Psi(1) - \Psi(2)]$$

$$= (n/b)[1 + \Psi(1) - \Psi(1) - 1] = 0 ,$$

$$E[\nabla_c] = -n[(b-1)/a]\Gamma(1-1/b) + n[b/a]\Gamma(2-1/b)$$

$$= -n[(b-1)/a]\Gamma(1-1/b) + n[b/a](1-1/b)\Gamma(1-1/b)$$

$$= 0 .$$

The symmetric information matrix,

$$\underline{M} = E[-\underline{H}(T)] = \{E[-h_{ij}]\} ,$$

can be derived from the terms of the symmetric Hessian and the Weibull density, $f_2(x,T)$, as follows:

$$E[-h_{11}] = (nb/a^2)[(b+1)\Gamma(2) - 1] = n(b/a)^2$$

$$E[-h_{12}] = (n/a)[1-\Gamma(2)-\Gamma'(2)] = -(n/a)\Gamma'(2) = -(n/a)\Gamma(2)\Psi(2)$$

$$= -(n/a)\Psi(2) = -(n/a)(0.42278\bullet\bullet\bullet)$$

$$E[-h_{13}] = (b/a)^2\Gamma(2-1/b) ,$$

$$E[-h_{22}] = n/b^2 + \Gamma''(2) = n/b^2 + [\Psi'(2) + \Psi^2(2)]\Gamma(2)$$

$$= n/b^2 + 0.46619\bullet\bullet\bullet$$

$$E[-h_{23}] = (n/a)[\Gamma(1-1/b) - \Gamma(2-1/b) - b\Gamma'(2-1/b)]$$

$$= (n/a)[\Gamma'(1-1/b) - \Gamma(2-1/b)\{1 - b\Psi(2-1/b)\}] ,$$

$$E[-h_{33}] = n[(b-1)/a^2][\Gamma(1-2/b) + b\Gamma(2-2/b)]$$

$$= n[(b-1)/a]^2\Gamma(1-2/b) .$$

Ravenis[198] gives the information matrix for the Weibull family, $f_1(x,T)$, and Harter and Moore[118] give similar numerical results for singly censored Weibull samples.

We recall that the inverse of the information matrix is the Cramer-Rao minimum variance bound discussed earlier. This inverse can be derived algebraically, but the usefullness of this explicit result does not warrant the space and effort required for derivation and display here. Although Huzurbazar[132] has shown that for any (multivariate) density of the Koopman-Pitman family -that is, any density admitting sufficient statistics- the M.L.E. asymptotically acnieve the bound, so that the inverse of the information matrix is the variance covariance matrix for $\hat{T}$, the full parametric Weibull family is not a Koopman-Pitman form. Fortunately, Halperin[113] generalizes the Cramer-Rao minimum variance bound result under mild regularity conditions to any density and also establishes asymptotic unbiasedness, consistency and normality for M.L.E.

For the Weibull family of densities, M.L. estimation is regular[118] for complete samples only if the location parameter, c, is known, or if the shape parameter, b, is greater than 2. We can verify above, in fact, that the term $E[-h_{33}]$ in the information matrix has a singularity for b=2,

but that all terms are well defined for b>2. Huber[131] and Le Cam[149] investigate the effect of nonregularity on M.L.E. properties.

Since for any regular parametric estimation the M.L.E. asymptotically achieve the Cramer-Rao bound given by the inverse of the information matrix, $E[-\underline{H}(\hat{T})]^{-1}$, and this inverse, a variance covariance matrix, is known to be positive definite[141,p.56], we may take some comfort in the knowledge that at least in expectation $\underline{H}(\hat{T})$ is negative definite. A stronger result follows immediately for simply consistent M.L.E., which must converge in probability to a single set of parameters. Thus, even for a problem with multiple solutions, the coordinates of the maxima must approach the same point asymptotically. Chandra[46] presents evidence for consistency of M.L.E. under very general conditions. We also recall that M.L.E. are asymptotically functions of sufficient statistics, when such exist. Huzurbazar[132] shows that sufficiency also implies uniqueness for $\hat{T}$.

Choi[48] gives estimates of bias for M.L.E. in the closely related Gamma family of densities, finding the magnitude of bias to be small even for intermediate sample sizes. Harter and Moore[117] suggest that a local maxima, though not a true M.L.E. in the strictest global sense, can exhibit most of the desirable statistical properties.

For finite sample sizes in cases such as the Weibull where no sufficiency can be established, the asymptotic results above do not necessarily hold. However, Harter and Moore[118] have performed extensive simulation studies for the three parameter Weibull family, and give tables showing that the Weibull M.L.E. achieve their asymptotic variances very quickly, with the actual variance exceeding the

Cramer-Rao bound by an amount proportional to $1/n^2$. Thus, their results suggest that the inverse of the information matrix is valid as an estimate of the M.L.E. variance covariance matrix for samples as small in size as 100. When a, or b, are known, this result is achieved much more rapidly.

As mentioned earlier, when the shape parameter, b, is less than 2, a nonregular estimation results, and if $b<1$, $L(X,T)$ is not bounded. Pike[185] suggests that when the likelihood function is unbounded in the feasible parameter space, a local maxima provides a reasonable estimate. It is easiest to constrain the range of the shape parameter to avoid this problem, since in a pure sense the M.L. method is inapplicable otherwise. Rockette, Antle and Klimko[202] suggest substitution of

$$T^+ = \{ a^+, b^+, c^+ \} ,$$

with

$$a^+ = \sum_{i=1}^{n} (x_i - c)/n ,$$

$$b^+ = 1 ,$$

and

$$c^+ = x_{[1]} ,$$

for cases in which no M.L. solution exists. For cases in which a maximum is achieved, it is compared with T^+, and the

M.L. solution $\hat{T}$ is chosen to correspond with the higher likelihood.

As a useful simplification of the three parameter Weibull estimation, the scale parameter can be eliminated from the system by solution of $\nabla_a = 0$ and substitution, which gives

$$\hat{a} = [\sum_{i=1}^{n} (x_i - c)^b / n]^{1/b} ,$$

so that

$$T^* = \{ \hat{a}, b, c \} ,$$

for which the conditioned log likelihood function becomes

$$L^*(X,T) = n \ln(b) - n \ln[\sum_{i=1}^{n} (x_i - c)^b / n]$$

$$+ (b-1) \sum_{i=1}^{n} \ln(x_i - c) - n ,$$

with gradient

$$\nabla_b^* = n/b - n[\sum_{i=1}^{n} (x_i - c)^b \ln(x_i - c) / \sum_{i=1}^{n} (x_i - c)^b]$$

$$+ \sum_{i=1}^{n} \ln(x_i - c) ,$$

$$\nabla_c^* = nb \sum_{i=1}^{n} (x_i - c)^{b-1} / \sum_{i=1}^{n} (x_i - c)^b - (b-1) \sum_{i=1}^{n} [1/(x_i - c)] .$$

The symmetric Hessian, $\underline{H}^*$, for this reduced problem is

$$h^{*}_{22} = - n/b^{2}$$

$$- n\sum_{i=1}^{n}(x_i - c)^{b}\ln^{2}(x_i - c) \Big/ \sum_{i=1}^{n}(x_i - c)^{b}$$

$$+ n\Big[\sum_{i=1}^{n}(x_i - c)^{b}\ln(x_i - c) \Big/ \sum_{i=1}^{n}(x_i - c)^{b}\Big]^{2},$$

$$h^{*}_{23} = n\{b\sum_{i=1}^{n}(x_i - c)^{b-1}\ln(x_i - c) + \sum_{i=1}^{n}(x_i - c)^{b-1}\}$$

$$\Big/\sum_{i=1}^{n}(x_i - c)^{b}$$

$$- nb\sum_{i=1}^{n}(x_i - c)^{b}\ln(x_i - c)\ \sum_{i=1}^{n}(x_i - c)^{b-1}\}$$

$$\Big/\Big[\sum_{i=1}^{n}(x_i - c)^{b}\Big]^{2}$$

$$-\sum_{i=1}^{n}(x_i - c)^{-1},$$

$$h^{*}_{33} = nb^{2}\Big[\Big\{\sum_{i=1}^{n}(x_i - c)^{b-1}\Big\}^{2}$$

$$- \sum_{i=1}^{n}(x_i - c)^{b}\sum_{i=1}^{n}(x_i - c)^{b-2}\Big] \Big/ \Big[\sum_{i=1}^{n}(x_i - c)^{b}\Big]^{2}$$

$$+ nb\sum_{i=1}^{n}(x_i - c)^{b-2} \Big/ \sum_{i=1}^{n}(x_i - c)^{b}$$

$$- (b-1)\sum_{i=1}^{n}[x_i - c]^{-2}.$$

The expectation of $\underline{H}^{*}$ is hopelessly difficult to derive, so that additional assertions of uniqueness, or other properties, are not achievable by that method for this reduced system. However, Peto and Lee[184] treat each element of the gradient as an implicit function of the other respective M.L. estimator, and plot the two trajectories for $\nabla_{T} L^{*}(\mathbf{X},\mathbf{T})=0$, showing the M.L. solution as an intersection.

Rockette[201] carries this treatment further, and obtains strong, but not conclusive evidence via lengthy term by term inequality arguments for the sums of powers of X found in the conditioned Hessian (drawing from Hardy, Littlewood and Polya[114]) that at most one maximum exists for the three parameter Weibull M.L. problem, and that there is a saddle point associated with each maximum. Extensive empirical evidence supports these assertions.

A generalization of the Weibull likelihood model is possible for other than complete samples. Cohen[52] classifies the process by which elements have been censored from a sample of failure times in life testing as being of Type I when the sampling process is stopped at some predetermined time, or Type II when testing ceases after some fixed number of elements, k, have failed. For Type I censoring, the number of observed failure times, k, is a random variable, while Type II censoring provides a random sampling cessation time. For either type of censoring, when only the first k out of n elements have been observed, the Weibull likelihood function is

$$L_k(\mathbf{X},T) = \ln\{[n!/(n-k)!] \prod_{i=1}^{k}[f_2(x_i,T)]\, R_2(x_k,T)^{n-k}\}$$

$$= \ln[n!/(n-k)!] + k[\ln(b) - b\ln(a)]$$

$$+ (b-1)\sum_{i=1}^{k}\ln(x_i - c) - \sum_{i=1}^{k}[(x_i - c)/a]^b$$

$$- (n-k)[(x_k - c)/a]^b$$

with x_k fixed for Type I censoring, and $x_k = x_{[k]}$ for Type II.

Ringer and Sprinkle[200] and Cohen[53] discuss M.L.E. for the censored Weibull density with c=0, and Harter and

Moore[11b] study the three parameter case for the Weibull and Gamma families.

Progressive censoring of either type occurs when at successive stages a number of survivors are randomly selected for removal from further testing. Suppose that s stages of censoring occur at progressive times y_j, $j=1,2,\ldots,s$, with k_j functioning elements randomly removed at each stage from further testing. Progressive Type I censoring, with the predetermined constant times, Y, produces a random sample size

$$m = n - \sum_{j=1}^{s} k_j ,$$

and has log likelihood

$$L^{I}_{Y,K}(\mathbf{X},\mathbf{T}) = \ln\{ C \prod_{i=1}^{m} f_2(x_i,T) \prod_{j=1}^{s} R_2(y_j,T)^{k(j)} \}$$

$$= \ln(C) + m \ln(b/a) + (b-1) \sum_{i=1}^{m} \ln[(x_i - c)/a]$$

$$- \sum_{i=1}^{m} [(x_i - c)/a]^b + \sum_{j=1}^{s} k_j [(y_j - c)/a]^b$$

with C a combinatorial constant.

For Type II progressively censored sampling, with the number of censored elements, k_j, fixed, and the times of censorship occurring randomly with each failure, the log likelihood is

$$L^{II}_{Y,K}(X,T) = \ln\{\prod_{i=1}^{s}[(n - \sum_{j=1}^{i-1}[k_j - i + 1])\ f_2(y_i,T)\ R_2(y_i,T)^{K(i)}]\}$$

$$= \sum_{i=1}^{s} \ln\{[(n - \sum_{j=1}^{i-1}[k_j - i + 1])$$

$$+ (b-1)\sum_{i=1}^{s}\ln[(y_i - c)/a]$$

$$- \sum_{i=1}^{s}[(y_i - c)/a]^b + \sum_{i=1}^{s} K_i[(y_i - c)/a]^b$$

The gradient and Hessian matrix for each of these models is not included here. It should be pointed out, however, that the scale parameter can be eliminated in each model in precisely the same manner that gave $L^{*}(X,T)$. Ringer and Sprinkle[200] give the gradient for $L^{II}_{Y,K}(X,T)$ when c=0, and Cohen[53] gives the gradient and Hessian for $f_1(X,T)$ with c=0. Wingo[234] gives the gradient and Hessian for Type I progressive censoring of $f_1(X,T)$.

Harter and Moore[118] give the gradient and Hessian for doubly censored, or truncated, samples in which the first k and last r elements have not been observed. The form of the log likelihood function is very similar to the singly censored case. An interesting result of these empirical studies of censored sampling is that when the location parameter, c, is not bounded from the left a higher variance results for $\hat{c}$ than when c is constrained

$$0 < c < x_{[1]}.$$

Rockette[201] reports that this effect seems to increase with the ratio, c/a. Polfeldt[188,187] has given some limited theoretical results for nonregular estimation of location parameters. Antle and Bain[ ] have given several interesting transformations of scale and location parameters which are statistically independent.

Note that a numerical singularity occurs when c approaches $x_{[1]}$ too closely in a complete sample. This suggests using $\hat{c}=x_{[1]}$ and dropping $x_{[1]}$ from the sample when necessary. For large samples this heuristic can be extended to the censoring of all sample elements in the close neighborhood of $x_{[1]}$. As a practical matter, this adjustment can circumven. serious difficulties with M.L. estimatic for some samples, but it is somewhat distasteful to peremptorily discard costly sampling information in this way. Also, strongly assymetric censoring can introduce more bias for the M.L.E., and of course increase M.S.E. Harter has reported that even for sample sizes of 10 and 20, bias is not severe under moderate censoring and the theoretical variance of the M.L.E. is not greatly exceeded.

The reluctance with which $\hat{T}$ approaches the Cramer-Rao bound for intermediate sample sizes can be overcome by constraining the shape parameter, b, to a feasible range known by the investigator. The M.S.E. can be lowered significantly by such precaution, since the Weibull density function and consequently the likelihood objective function are very unruly for high values of b, unbounded for b=1, and nonregular for b<2. If we constrain b to values between, for instance, 1 and 4 we have still included a very robust parametric family in our investigation, but one with less habitual inclination to provide ridiculous likelihood estimates for purely numerical reasons.

For very small samples, an investigator is faced with the unfortunate paradox that, although the objective function and its derivatives are easily and quickly evaluated, the likelihood surface can exhibit a tortuous landscape. Perhaps this is fortuitous, for otherwise one would be tempted to rely on $\hat{T}$ despite its unknown statistical properties. The irregularity introduced by including the location parameter, c, in the search is most troublesome for these cases. The frequent occurrence of a stationary saddle point usually takes place at parametric coordinates relatively close to the upper bound for c, $x_{[1]}$; however the saddle point can lie well within the range of c for small samples, making it difficult to consistently identify and avoid numerically.

If a sample is used for the M.L. estimation with the Weibull model that actually comes from some markedly different population, the results can be disastrous even for large sample sizes. It is worthwhile to remember that the statistical theory underlying this estimation process requires that the hypothetical assumption of the density function for which point estimates are sought must be based in fact. Two singular examples of such (large) samples have come to the author's belated attention in this regard; one was subsequently identified as coming from a pareto population, and the other was ultimately determined to be a sample from a beta density. These samples wreaked numerical havoc with several optimization codes applied to the Weibull model, the first because of too many sample elements in the extreme right tail for the Weibull density to fit, and the second due to the effect of a finite upper domain limit for a symmetric sample. Both samples produced apparent stationary saddle points, numerically unbounded $\hat{T}$, and infinite likelihoods at various times. Thus, great care must be taken in applying any numerical M.L. procedure to

the Weibull, since termin[illegible]t a stationary point on L(X,T) should be allowed only for a maximum, or the optimization problem should terminate with indication of no achievable finite optimum.

Inferential techniques based on finite samples for Weibull and other closely related models are given by Harter and Moore[115], Bain and Weeks[9], Thoman, Bain and Antle[221], Bain[8] and Billman, Antle and Bain[23]. Most of these investigations give tables which are developed by estensive simulation.

M.L. estimation of the reliability function is shcwn to be surprisingly unbiased and robust by Hager, Bain and Antle[112] and others.

C. NUMERICAL APPROACH

The methods commonly used for numerically determining Weibull M.L.E. have historically included cyclic search by Harter and Moore[116], second order Newton-Raphson ascent by Peto and Lee[174] and Ringer and Sprinkle[200], and quasilinearization (Newton-Raphson ascent with the Hessian approximated by differences) by Wingo[233,234].

The magnitude of effort involved in applying a search technique to M.L. estimation is implied by the iteration limit suggested by Harter and Moore of 550 cycles for the Weibull model and 1100 cycles for a Gamma. Barnette[14] suggests cyclic search for likelihood models with multiple roots; such cases usually occur only for small samples from selected density functions. The Weibull model has never empirically exhibited finite multiple optima, although some small sample estimations lead to numerical difficulties characterized by a stalling of the iteration over a respectable neighborhood.

In a refreshingly honestly titled article, Mantel and Myers[162] report that for second order ascent methods choice of the starting value, T_0, is vital to success. (This is, of course, no theoretical surprise to a numerical analyst.) As we have seen, the Weibull model seems to produce a saddle point as a gratuitous companion of a maximum. For this reason, pure second order representations of the problem have consistently not lead to acceptable performance of the optimization algorithm.

Kale[137,138] compares the second order Newton-Raphson and Fisher's scoring methods and indicates that they are most applicable for large samples. Unfortunately, both his

iteration and residue criteria for rate of convergence evaluation are not directly related to computer time, and his sample problem is a fairly uncomplicated two parameter estimation. Michelini[167] gives a method for selecting starting values for the scoring method applied to a lognormal model, and presents fascinating graphical depiction of empirical regions of convergence.

Implementation of both first and second order ascent methods and search techniques for the Weibull models have produced the following conclusions. The first order gradient methods are superior to search techniques for reasons of speed, and are better than second order iteration on the basis of reliable convergence. The saddle point dilemma is most expeditiosly resolved by use of first order methods and a solution verification; convergence to a saddle point is very rare for the constrained parameter problem and sample sizes greater than 10.

All techniques regularly fail for small samples. It is suggested that for these cases either the sample has insufficient information to warrant M.L.E., or the wrong density is being used for parametric estimation.

A hybrid ascent method which produces both fast and dependable convergence for highly nonlinear problems utilizes both first and second order representations of the maximization problem. The first order formulation is used to begin the solution, and continued until the amount of information in the linear term of the objective function approximation diminishes significantly below the remaining higher order terms in

$$L(T_0 + \Delta T) = L(T_0) + \nabla L(T_0)'\Delta T + rem(T_0, \Delta T)$$

The switching rule requires that a sequence of solutions of specified length exhibit a moving average of

$$\text{avg}\{\ L(T_0)\cdot\Delta T\ /\ \text{rem}(T_0,\Delta T)\} \leq B_4 .$$

B_4 has been set at 0.1. The transition to the second order representation is not guaranteed for every problem, since it occurs only when higher order terms dominate the local approximation of L.

To test these methods, 50 samples of size 100 were randomly generated from Weibull families with (arbitrarily) a=50, c=100, and b=1.5,2.0,2.5,3.0. For all estimations a total computation time was recorded with the iteration records. The host computer was an IBM 360/67 - II with optimized FORTRAN-IV(H). A pure first order representation, and the hybrid scheme were used on the matched set of random samples, with B_4=0.1 and the length of the moving average for the switching rule set at two.

The results were

			AVERAGE PERFORMANCE				
n	b	METHOD	a	b	c	Time(sec)	ITERATIONS
100	1.5	I	47.95	1.56	101.28	46.0	59.3
			4.47	.19	1.67		
		II	48.01	1.52	101.27	18.2	21.3(12.7)
			4.94	.17	1.55		
	2.0	I	48.56	1.98	101.19	50.9	59.2
			5.00	.30	3.28		
		II	48.54	1.97	101.19	20.9	19.6(4.8)
			4.98	.30	3.29		
	2.5	I	49.23	2.49	100.83	45.7	51.8
			5.26	.37	4.92		
		II	49.25	2.49	100.80	19.1	17.0(3.9)
			5.22	.36	4.90		
	3.0	I	47.80	2.97	101.75	51.4	67.6
			6.69	.56	6.21		
		II	47.76	2.99	101.72	16.6	15.9(4.1)
			6.64	.55	6.18		

The root mean squared error is given just below the sample mean for estimators of each parameter. The number in parentheses following the average iterations required for convergence of the second order scheme indicates the average number of first order iterations required to trigger the switching rule. There were no cases for which convergence was not achieved. The second order representation clearly dominates these results.

For b=1.5, seven cases converged with results which were replaced by the solution, T^{+}, (with b=1) on the basis of likelihood comparison. Samples generated for other higher values of b produced no such replacements. For samples from the population with b=3.0, six cases converged to the upper numerical bound, b=4, and did not achieve transition to the second order representation. Since this

seemed to have no serious effect upon the expectation of b for this set of samples, no further action was taken. Raising the bound for b might be indicated in other investigation contexts.

All estimations were started with the initial solution b=2 and c=0.9 $x_{[1]}$. In order to test the robustness of the numerical methods with respect to starting values, the same sets of random samples were used with initial values of b=1.5,2.0,3.0. The numerical values of the M.L.E. results for each sample showed almost no sensitivity to starting value, and the average number of iterations required for convergence and the computation time consumed were not significantly different, although individual samples did occasionally exhibit large variations. For several samples, the first order method converged to values differing in the hundredths position. Such variation was never evident for the second order hybrid iteration.

The behavior of the objective function during optimization with the hybrid second order method was remarkably consistent for all samples. At the initial iteration, the first order term in the linear approximation of the objective function dominated the remainder term. After several iterations, as indicated in the performance data, transition to the second order representation of the problem took place, after which no more than three iterations produced a solution for which the linear approximation of the second order representation objective function left a remainder term several orders of magnitude smaller than the linear term. The final likelihood achieved by the second order scheme was higher than that given by the pure first order method in every case, but the difference

never exceeded a relative magnitude of 10^{-5}.

By "tuning" the iteration schemes, significant further reductions in computing time are possible. This is especially true if the resolution of the termination criteria is relaxed. This is usually a reasonable course of action, since the values of the M.L.E. are not really required to, say, five decimal places. Such a change for the second order model with b=2 produced an average convergence time of 9.7 seconds in 11.3(3.4) iterations.

It is important to note that for various sample sizes, termination has always been achieved without numerical processor interrupts. That is, the algorithm detects terminal singularities, and indicates them under full control of the program, permitting remedial action to be taken, or simply allowing analysis of the complete output. This is far superior to the spectacular results to be expected from most Newton-Raphson based programs.

There is no theoretical reason prohibiting formulations of even higher order representations of nonlinear optimization problems. The motivation for such an approach would be a highly nonlinear problem for which the first order approximation of the imbedded local linear program can be expected to give a poor representation of functions, even with the second order formulation. The algebraic demands of higher order representations of likelihood models and the consequent debugging and computational expense of the associated higher order problems do not promise much practical value. Fortunately, the likelihood models investigated thusfar have not reached such nonlinear extremes within the numerical capabilities of a digital computer, as is attested by the dominated remainder terms for the second order representation. However, it is

possible that for other types of highly nonlinear problems high order representations will prove a fruitful field for further research. A sequential transition mechanism such as that proposed here may also provide for robust convergence with higher order formulations as it has done in the present investigation.

D. A NONLINEARLY CONSTRAINED PROBLEM

Consider the following problem. A random sample is collected from a Weibull population whose mean is known, but whose parameters are to be estimated by M.L. Such situations arise, for instance, when census information is used in conjunction with random survey data for demographic modelling.

As another example, suppose a bank wishes to use a random sample to estimate the parameters of a Weibull density function describing the size of an individual depositor's account. Certainly the bank will know exactly the total of money on deposit and the number of depositors. Thus the mean of the density is known, but not the parameters.

The M.L. formulation of such a constrained problem becomes

$$\begin{aligned} &\underset{T}{\text{MAX}} \quad L(X,T) \\ \text{s.t.} \quad &c + a\Gamma(1+1/b) = m_1 , \\ &0 < a < \infty , \\ &1 < b < 4 , \\ &0 < c < x_{[1]} . \end{aligned}$$

The constraint has gradient elements

$$g_a = \Gamma(1+1/b) ,$$

$$g_b = -(a/b^2)\Gamma'(1+1/b)$$

$$= -(a/b^2)\Gamma(1+1/b)\Psi(1+1/b) \ ,$$

$$g_c = 1 \ ,$$

and the Hessian for the constraint, $\overset{\#}{\underline{H}}$, has the nonzero terms:

$$\overset{\#}{h}_{12} = -1/b^2\Gamma'(1+1/b)$$

$$= -1/b^2\Gamma(1+1/b)\Psi(1+1/b) \ ,$$

$$\overset{\#}{h}_{22} = (a/b^3)\Gamma'(1+1/b) + (a/b^4)\Gamma''(1+1/b)$$

$$= (a/b^4)\Gamma(1+1/b)[\,b\Psi(1+1/b) + \Psi'(1+1/b) + \Psi^2(1+1/b)\,] \ .$$

As before, the scale parameter, a, may be substituted out, leaving

$$\underset{T}{MAX} \quad L^*(X,T)$$

$$\text{s.t.} \quad c + \left[\sum_{i=1}^{I_A} (x_i - c)^b/n\right]^{1/b}\Gamma(1+1/b) = m_1 \ ,$$

$$1 < b < 4 \ ,$$

$$0 < c < x_{[1]} \ .$$

The gradient for this reduced problem is

$$g_b^* = \Big[\sum_{i=1}^{n}(x_i - c)^b/n\Big]^{1/b}\Gamma(1+1/b)$$

$$\{(1/b)\Big[\sum_{i=1}^{n}(x_i - c)^b \ln(x_i - c)\Big/\sum_{i=1}^{n}(x_i - c)^b$$

$$-(1/b)\ln\Big(\sum_{i=1}^{n}(x_i - c)^b/n\Big)\Big]$$

$$+\Psi(1+1/b)\}\ ,$$

$$g_c^* = -\Gamma(1+1/b)\Big[\sum_{i=1}^{n}(x_i - c)^b/n\Big]^{1/b-1}\Big[\sum_{i=1}^{n}(x_i - c)^{b-1}/n\Big]\ .$$

The Hessian for the constraint will not be given here.

As a test of this model, ten samples of size 100 were randomly generated with a=50, b=2, and c=100, and the constrained mean, m_1, was set at

$$m_1 = c + a\Gamma(1+1/b) = 100 + 50\Gamma(1.5) = 144.31\bullet\bullet\bullet\ .$$

The three variable model was run for both first order and hybrid schemes, with the switching rule qualified to activate for feasible solutions only. The results were

			AVERAGE PERFORMANCE				
n	b	METHOD	a	b	c	Time(sec)	ITERATIONS
100	2.0	I	48.40	1.98	101.21	132.8	135.4
			4.65	.38	3.95		
		II	49.11	2.00	101.27	89.7	87.2(6.4)
			4.69	.39	4.02		

One sample did not make a transition to the second order representation before convergence, and two of the samples required 200 iterations for termination.

The choice of width for the equation band representing the equality constraint feasibility region and the manner in which this band is closed during the progress of the solution is most important for insuring success. Premature closing, or choice of a band too narrow can cause the methods to stall, especially for the second order representation which has polygamma terms that are exceedingly difficult to compute precisely. On the other hand, too wide a band, or too much delay of the closing can lead to excessive iterations involving infeasible solutions. The choice of a bandwidth of 0.1 was made in these applications with good success and the band was closed by 32 successive bisections for feasible solutions. Upon transition to the second order representation of the problem, it was determined that a reinitialization of the equation band had a desirable effect on convergence.

The superiority of the second order representation of the constrained model would be enhansed greatly by the use of efficient and/or accurate polygamma functions. Currently, the best series approximations derived produce only six decimal place precision, and their computation requires almost half of the iteration time reported above.

Other side constraints can be added to parametric M.L. estimation. For instance prior knowledge of the population variance, or other moments, can be used in the estimation. The numerical details follow directly from the example given here.

The results for both classical unconstrained and constrained models reported here for the parametric Weibull family apply with remarkably little modification to the general gamma family of densities as well. Regularity conditions, gradients, Hessians, numerical approach and so forth follow the Weibull examples very closely.

CHAPTER IV

A STRUCTURAL MODEL: BERNOULLI REGRESSION

A. INTRODUCTION TO A BERNOULLI REGRESSION MODEL

Consider a structural model based on the observed sample

$$\{Y , \underline{X}\} ,$$

where y_j is one of a set of n (statistically) independent discrete-valued observations with m associated parameters

$$X_j = \{ x_{j1}, \ldots, x_{jm} \} .$$

In this usage, $\underline{X}$ is often called the set of (structurally) independent variables, and Y is referred to as the associated (structurally) dependent variable, with T a set of model parameters.

As a specific example, suppose that Y is a set of "1-0" observations of "success, or failure" from n Bernoulli trials. We may assert that y_j is observed with

$$f(y_j|X_j,p) = f(y_j|p(X_j)) = f(y_j|p_j) ,$$

and that f is a parametric family of Bernoulli densities

$$f(y_j|p_j) = p_j^{y(j)} (1-p_j)^{1-y(j)} ;$$

$$0 \leq p_j \leq 1 \; ; \quad y_j = 0, \; 1 \; .$$

In the regression, p_j is some stated mathematical function of the independent variables, X_j, with parameters, T,

$$p_j = p(X_j, T) \; ,$$

and is interpreted as the prior probability of success for a Bernoulli trial carried out under a given set of conditions.

To illustrate, suppose that p_j is the probability of destroying, or disabling, a target with a volley of shots in a naval bombardment. The success of each volley can be considered as an observation, y_j, from a Bernoulli density with parameter p_j. Clearly, the probability of success on each attempt is a function of distance to target, sea conditions, weapons employed, visibility, and so forth - characteristics which constitute X_j.

If we employ the theory of ballistics to determine a functional form for $p(X_j)$, and if a record is kept by the fire director of each volley, then we have just the observations required to estimate p_j.

Consider another example. Let p_j be the probability of

a smog alert during a particular day. If a record of wind velocity, wind direction, temperature, nitrous oxide level, cloud cover, particulate content, and so forth, is kept daily constituting X_j, with y_j the observation of a smog alert for that day, then estimation of p_j may be attempted from n independent observations of polluted and unpolluted days, with some function, $p(X_j)$, supplied by the researcher.

The Bernoulli parameter, p_j, may be the probability of default on a loan given credit information X_j the probability of winning an election given a platform and legislation record, the probability of survival given information about disease and treatment, ad infinitum.

It should be stressed that discrete Bernoulli observations are often available when continuous quantitative information is not, or when continuous measure is inappropriate. For instance, it may be possible to classify an individual as "poor" while to use a measure of his economic income would be difficult or impossible due to unreported income, government subsidy in the form of money, goods and services, unclear family consuming units, and the problematic equivalence of income level with the quality of life.

As another illustration, the regression analysis of a communications satellite launching may, for purposes of research budget request, properly deal with the probability of successful orbital entry, or launch failure, rather than with orbital apogee, perigee, period, etc. Thus all the information concerning launch conditions and technology would be used to yield a prior probability of success, a

result more tractable for management and more closely related to project costs than estimates of orbital physics.

It is felt that the general class of problems dealt with here is important and previously overlooked, or misclassified in the literature. Several Bernoulli models are presented in the sections that follow. Point and interval estimates for p_j are developed, a hypothesis test is given for evaluating the contribution of parameters to complicated, realistic models, a stepwise construction technique is proposed, and a heuristic is given for choosing between functional forms for the regression.

B. COMPARISON WITH DISCRIMINANT ANALYSIS

A technique often misapplied to Bernoulli regression problems is that of discriminant analysis. Borrowing prior notation in this new context, a binary discriminant analysis provides a decision rule for classifying an individual as a member of one of two populations (π_1, π_2) from examination of a set of k properties, X_j. Each individual is asserted to be a permanent member of only one of the populations. The discriminant analysis attempts to determine which of these mutually exclusive populations contains the individual.

For example, the Internal Revenue Service in this country uses a property set X_j consisting of income level, deduction types and quantities, etc., in order to classify an individual filing an income tax return either as a member of the population of chiselers, or honest tax payers. Those classified in the former population are audited in detail for errors and misrepresentations.

Applications occur frequently in the literature, and classically have included taxonomic classification by physical measurement, qualitative biochemical analysis, pattern recognition, identification of archeological remnants, and so forth. For excellent examples see Fisher[85] and Nilson[176].

The discriminant analysis requires use of n_1 known members of π_1, and n_2 individuals from π_2, and a density function for the property set of each population, $f_1(X)$ and

$f_2(X)$.

The probability of an observation in the neighborhood of the point X_j, given that the individual is from π_1, is

$$f_1(X_j)\ dX_j\ .$$

This probability is proportional to the argument $f_1(X_j)$ which is defined as the likelihood function for the point X_j. The fundamental principle of discriminant analysis is to classify the individual as a member of π_1, or π_2, according to the relative size of $f_1(X_j)$ and $f_2(X_j)$, and the costs of each type of misclassification. In general, the density functions f_1 and f_2 will contain unknown parameter vectors T_1 and T_2, and these parameters must be estimated from the known members of each population. The parametric estimation is usually performed with M.L.E., as previously discussed.

The discriminant analysis will further require the prior probability of selecting a member of π_1 for analysis,

$$Pr_1 = n_1 / (n_1 + n_2)\ ,$$

or, when population sizes are unknown, a sampling estimate of Pr_1 may be used. If no sampling information is available, and the population sizes are unknown, Pr_1 is

assumed to be 0.5.

Also, the costs of misclassification, $C_{2|1}$ and $C_{1|2}$, must be stated, where $C_{2|1}$ is the cost of misclassifying π_2 when the individual is actually from π_1. Without loss of generality, $C_{1|1}$ and $C_{2|2}$, the costs of correct classification, are taken to be zero.

Finally, the decision rule for discrimination is: classify X_j as a member of π_1 if

$$Pr_1 C_{2|1} f_1(X_j) > (1 - Pr_1) C_{1|2} f_2(X_j) ,$$

and classify X_j as a member of π_2 otherwise. This minimizes the expected cost of misclassification.

Clearly, the results above may be generalized to any number of populations. Such multiple discriminant analysis is required in machines for character recognition in which a hardware automaton carries out the analysis automatically in a fascinating way[175]. Another example of the technique is multipnasic screening of school children for physical and mental defects by tests and inexpensive profile measurements. In this manner, a single property set is examined in order to classify an individual as healthy, or medically defective in any of several ways. It is assumed that these defects, once identified, may be verified with certainty by a more thorough, and expensive examination.

In contrast to the discriminant, the Bernoulli regression model is not concerned with classifying

observations into permanent populations of success and failure, but rather with forecasting the probability, p_j, of an individual achieving a success. The implication is that repeated trials on the same individual will produce some successes, and some failures, and that the properties X_j are not uniquely those of a member of some population of successes, or failures. It is interesting to note that many applications of discriminant analysis in the literature are specious for just this reason.

C. MATHEMATICAL PRELIMINARIES

To proceed with Bernoulli regression one must choose a functional form for p_j in the Bernoulli density, and then estimate any unknown parameters, T, in this function using the observations

$$\{Y , \underline{X}\} ,$$

and remembering that Bernoulli regression must produce predictions satisfying

$$0 \leq p_j \leq 1 .$$

Among the mathematical transformations available for our use are a general linear model with

$$p_j = X_j T ,$$

$$0 \leq X_j T \leq 1 ;$$

an exponential

$$p_j = \exp\{-X_j T\} ,$$

$$0 \leq X_j T \leq \infty ;$$

anotner exponential

$$p_j = \exp\{-[X_j T]^2\} ,$$

$$0 \leq X_j T \leq \infty ;$$

the logistic function

$$p_j = [1 + \exp\{-X_j T\}]^{-1} ;$$

Urban's tranformation

$$p_j = 1/2 + \pi^{-1} \tan^{-1}\{X_j T\} ;$$

a trigonometric model

$$p_j = (1/2)[1 + \sin\{X_j T\}] ,$$

$$-\pi/2 \leq X_j T \leq \pi/2 ;$$

and so forth.

There are, of course, an infinite number of candidates, as in any regression problem. We have chosen each of these to contain the linear form $X_j T$. In this way, there is a single parameter in T associated with each characteristic in $\underline{X}$. This permits definition of $x_{j1}=1$ so that t_1 may be interpreted as an "intercept" parameter in each model. Also, addition and deletion of characteristics in $\underline{X}$ may be performed easily; this facilitates, for instance, introduction of additional variables to $\underline{X}$, to allow for nonlinear interaction of characteristics. Just as in least

squares regression, it is perfectly admissible, and useful, for the independent variables to take on discrete values (e.g., 0,1). Finally, we shall discover a salutary distributional property of a wide class of such models, and we will develop a method for comparing the efficacy of two, or more, models in any particular problem.

Note that several of the models given require a constraint on the linear form $\mathbf{X}_j\mathbf{T}$. This follows from the "1-0" constraint on p_j and the desirability of providing for p_j to be stated as a single valued function of the argument $\mathbf{X}_j\mathbf{T}$. Although such constraints can be accomodated numerically, their number grows directly with the number of observations. For ease of exposition, we choose an unconstrained model for complete development here. That is, the transformation used mathematically guarantees a feasible probability, p_j.

As our example, we will use M.L. estimation for the logistic transformation. Berkson[20] suggests the logistic function for bio-assay models. Also see the presentation given by Finney[84]. A development of L.S. estimation for a similar logistic model is given by Walker and Duncan[226]. We remember, though, that the L.S. assumptions do not lead to tractable distributional results, while the M.L.E. approach will yield excellent large sample properties with invariance.

The log likelihood for the parametric Bernoulli family is

$$L(P) = \sum_{j=1}^{n} [\, y_j \ln(p_j) + (1 - y_j) \ln(1 - p_j) \,] .$$

Since

$$\ln f(p_j) / p_j = y_j/p_j - (1 - y_j)/(1 - p_j) ,$$

we see that regardless of the parametric form for p_j, $E[L'(P)]=0$ by inspection. Parameterization with the logistic function

$$p_j = [1 + \exp\{-X_j T\}]^{-1} ,$$

gives

$$\ln(p_j) = -\ln(1 + \exp\{-X_j T\}) ,$$

and

$$\ln(1 - p_j) = -X_j T - \ln(1 + \exp\{-X_j T\}) .$$

From this we find the gradient

$$\nabla_i = \sum_{j=1}^{n} ([y_j x_{ji} \exp\{-X_j T\} - (1 - y_j) x_{ji}] / [1 + \exp\{-X_j T\}]) ,$$

and symmetric Hessian matrix

$$\underline{H}(T) = \{h_{ik}\} = \sum_{j=1}^{n} (-x_{ji} x_{jk} \exp\{-X_j T\} / [1 + \exp\{-X_j T\}]^2) .$$

The symmetric information matrix, $E[-\underline{H}(T)]$, is

$$E[-h_{ik}] = \sum_{j=1}^{n} (x_{ji} x_{jk} \exp\{-X_j T\} / [1 + \exp\{-X_j T\}]^2) .$$

(Remarkably, from this we see that the Newton-Raphson and Fisher's scoring methods are identical for this model.)

Once an M.L.E. solution, $\hat{T}$, has been determined for a particular problem, the invariance property gives for the single valued function p_j

$$\hat{p}_j(T) = p_j(\hat{T}) ,$$

and the Maximum Likelihood Estimation of the Bernoulli regression is complete.

D. STATISTICAL THEORY

We can derive confidence limits for the parameter p_j by noting that the M.L. estimators, $\hat{T}$, are asymptotically normally distributed with a variance covariance matrix given by the Cramer-Rao bound, the inverse of the information matrix,

$$\underline{V}(\hat{T}) = E[-\underline{H}(\hat{T})]^{-1} .$$

Using the logistic model as an example, clearly the scalar,

$$X_j \hat{T} ,$$

is a linear combination of asymptotically normally distributed random variables, and thus is asymptotically normally distributed with parameters

$$m_{1(j)} = X_j T ,$$

and

$$m_{2(j)} = X_j \underline{V} X'_j .$$

Confidence limits for $X_j T$ are

$$m_{1(j)} \pm z_{a/2} m_{2(j)}^{1/2} ,$$

where $z_{a/2}$ is the appropriate unit normal variate for

confidence level 1-a. Confidence limits for p_j follow directly by application of the logistic function to the limits for $X_j T$.

All the mathematical transformations, $p_j(X_j,T)$, proposed earlier for possible use in Bernoulli regression were chosen to be single valued functions of the scalar argument $X_j T$ to provide similar results in general.

To test the contribution of a particular parameter, or set of parameters, $\stackrel{?}{T}$, in the prediction of p_j, a likelihood ratio hypothesis test developed by Neyman and Pearson[173], Wald[224] and given in Mood and Graybill[169], can be used. As an example, to compare $p_j(T)$ and $p_j(\stackrel{*}{T})$, with

$$\stackrel{*}{T} = \{T,\stackrel{?}{T}\} ,$$

obtain $\stackrel{\wedge}{T}$ and $\stackrel{\wedge}{\stackrel{*}{T}}$, and compute the statistic

$$-2 \ln[L(\stackrel{\wedge}{T}) / L(\stackrel{\wedge}{\stackrel{*}{T}})] .$$

As n approaches infinity, the statistic is asymptotically chi-square with degrees of freedom equal to the number of elements in $\stackrel{?}{T}$, providing sufficient statistical grounds to give a constructive hypothesis test.

A stepwise Bernoulli regression algorithm can be

defined and performed by introducing each of the variables singly with an intercept term, determining the M.L.E. in each case, and keeping the characteristic leading to the greatest likelihood among the competing two-variable models. Sequential steps proceed, selecting one additional variable at a time by the criterion of maximum likelihood contribution among all remaining individual candidates. The procedure stops either when all variables have been included, or when addition of another variable produces a likelihood ratio less than the value of the chi-square integral for one degree of freedom and, say, 95 percent confidence (3.841•••).

To choose between alternative functional forms for the regression model, for instance between the logistic model and the Urban transformation, the following heuristic is proposed: perform the stepwise algorithm for both mathematical functions, and select the function for which the final likelihood is larger.

E. AN EXAMPLE: PREDICTION OF LABOR FORCE PARTICIPATION

An interesting problem to which Bernoulli regression is applicable is given by Solberg[214], who investigates the propensity of female heads-of-household with dependent children to join the labor force, and gives extensive references to and criticisms of published analyses using other statistical techniques.

The data used for analysis is extracted as a subset of the March, 1970, Person-Family file of the Current Population Survey conducted by the United States Census Bureau; the data selection criteria produces 2,839 observations of female family heads with dependent children present who are in the civilian non-institutionalized population with a primary source of income not gained from self-employment in agriculture.

In order to study only individuals who can reasonably be expected to participate in the labor force with non-zero probability, family heads over 70 years of age are deleted along with certain identifiable incomplete entries, leaving 2,222 observations.

Observation j is coded $y_j=1$ if the head-of-household is working, with a job but not working, or looking for employment, and $y_j=0$ for those heads who are at home, in school, unable to work, or have other reasons for not participating.

Sixteen independent variables are defined for each observation as follows

x_1 = 1, an intercept term,

x_2 = earned income in hundreds of dollars,

x_3 = welfare income in hundreds,

x_4 = unemployment compensation,

x_5 = other income,

x_6, x_7, x_8 = "0-1" regional codes for North-Central, South-Central and Western States respectively,

x_9 = "0-1" SMSA Labor Market Control Variable, equal to one for families located in SMSA Central Cities,

x_{10} = number of children present,

x_{11} = number of children present under six years of age,

x_{12} = number of other adults present,

x_{13} = "0-1" race code, equal to one for black head-of-household,

x_{14} = age of head-of-household,

x_{15} = age squared,

x_{16} = years of education.

A program was written in FORTRAN to access the observations on a mass storage device, providing features to select any desired subset of the observations, scale, or normalize the variables, selectively list observations, and obtain the M.L.E., $\hat{T}$, for either the logistic or Urban's transformation for $p_j(T)$, using a second order representation of the problem and numerically bounded variables. A high resolution timer provides active compute

time statistics for the host computer, an IBM 360/67 - II operated under the MVT system. The object program was generated by the FORTRAN-IV(H) compiler with code optimization, requiring a memory region of approximately 200K bytes.

Other program features include an automatic stepwise introduction of variables to a given minimal fixed model from remaining indicated candidates, with sequential selection made on the basis of maximum log likelihood contribution, and termination triggered by a likelihood ratio hypothesis test successively performed at each step with a level of significance specified by the user.

Also, a variance covariance matrix is given for any designated model solution, $\hat{T}$, by use of the inverse Hessian Cramer-Rao bound, and used to compute confidence intervals for $p_j(T)$ for specified observations in the original data set, or other source. The final regression model is applied to the data and a frequency distribution is individually produced for observations with $y_j=0$ and with $y_j=1$.

In our analysis, the logistic and Urban's transformations were separately applied both stepwise and simultaneously to all sixteen variables, a ten variable subset consisting of

$$\{Y, X_i,\ i=1,2,3,4,5,9,11,13,14,16\}\ ,$$

and an eight variable subset comprised of

$$\{Y, X_i,\ i=1,2,3,5,11,13,14,16\}\ .$$

On the basis of the log likelihood heuristic proposed earlier, and on the apparent reluctance with which Urban's transformation produces predicted probabilities near zero, the logistic model was selected for further detailed analyses.

As an example of the results, the eight variable composite gave a stepwise logistic model with variables introduced in the sequence indicated with each step:

STEP	1	3	2	16	5	11	14	13
1	1.111	-.042						
2	.162	-.016	.300					
3	-.025	-.016	.298	.011				
4	.206	-.019	.292	.011	-.507			
5*	.399	-.018	.289	.013	-.643	-.539		
6	.945	-.019	.287	.012	-.571	-.690	-.012	
7	.530	-.020	.286	.013	-.545	-.698	-.012	.036

The final log likelihood for this model is -795.2. The asterisk indicates the six variable model for which a 95 percent likelihood ratio test, with critical chi square value 3.841, would terminate with log likelihood -797.9.

Execution time for this run includes disk access, M.L. estimation, comparison and output of seven two-variable models, each with 2,222 observations, six three-variable models, five four-variable models, and so forth, yielding an aggregate to 10 minutes, 14 seconds.

For specific subproblems, an individual M.L. estimation is not permitted by the program to require more than ten iterations. This bound was never exercised by the Bernoulli models discussed here. The eight variable model required an average at all levels of search dimensionality of 4.1

iterations for convergence.

It is important to note the remarkable step to step stability of individual terms in $\hat{T}$, with the exception of the intercept term. This clearly shows that use of the previous solution as a starting value for successive iterations can greatly accelerate convergence of the second order representation of the problem. Exploitation of such behavior in nonlinear estimation has been suggested by Ross[207].

The regression predictions for p given for the 2,222 observations by the final logistic model are given in the following frequency distribution

FORECAST	ACTUAL	
p	y=0	y=1
0-.1	139	5
.1-.2	308	37
.2-.3	232	54
.3-.4	134	60
.4-.5	35	65
.5-.6	24	61
.6-.7	9	51
.7-.8	11	62
.8-.9	12	117
.9-1	34	772

In another experiment, a subset of 400 observations was randomly selected and $\hat{T}$ determined without stepwise introduction of variables for the sixteen variable logistic model. The solution of this pilot model was then used as a starting value for computation of $\hat{T}$ for all 2,222 observations, with a total computation time of 3 minutes, 21

seconds. A direct estimation without this preliminary step required 6 minutes, 8 seconds. Constructive stepwise estimation of all sixteen variables with no pilot models required 80 minutes, 42 seconds.

Specification of the appropriate size of such a pilot run is difficult, since a subset too small will give a solution of doubtful value (numerically and statistically) for starting the larger model, and a subset too large defeats the purpose of the approach. It is such small sample cases that exercise the numerical bounds and other provisions for difficulties in estimation. As a rule of thumb, 400 Bernoulli observations are used here with good success for "all at once" models.

Experience with all these models indicates that the constructive stepwise approach, possibly begun with a minimum model based on the investigator's prior experience, and terminated by the likelihood ratio test, is a generally reasonable plan of attack. Although computation time can be very high with such a method, important benefits are derived from model analysis by the M.L. estimation of subset models in the course of solution. For instance, subtle concomitance among variables may be detected by analysis of intermediate output that would evade detection in a final variance covariance matrix.

Other Bernoulli regression models have been studied, including prediction of the probability of winning a horserace, based on handicap data, and the estimation of the probability of increase in stock price from market analysis and financial information in investment survey guides. Most recently, DeMont and White[70] report analysis of tactical data from tank engagements in the Arab-Israel conflicts.

Generalization of these techniques is possible to

accomodate binomial, or multinomial data and models with dependent observations, although much work remains to be done.

BIBLIOGRAPHY

(1) Avadie, J. Integer and Nonlinear Programming, Elsevier, New York, 1970.

(2) Aitchison, J. and Silvey, S. D., "Maximum-Likelihood Estimation Procedures and Associated Tests of Significance," Journal of the Royal Statistical Society, B, 22, 1960, p.154.

(3) Amor, J. P., "Acceleration of Methods of Steepest Descent via Optimal Scaling," PHD Dissertation in Management, University of California, Los Angeles, 1973.

(4) Anderson, T. W. An Introduction to Multivariate Statistical Analysis , Wiley, New York, 1958.

(5) Antle, C. E. and Bain, L. J., "A Property of Maximum Likelihood Estimators of Location and Scale Parameters," Society for Industrial and Applied Mathematics, Review, 11, 1969, p.251.

(6) Arrow, K. J. and Enthoven, A. C., "Quasi-Concave Programming," Econometrica, 29, 1961, p.779.

(7) Arrow, K. J., Hurowicz, L. and Uzawa, H., "Constraint Qualification in Maximization Problems," Naval Research Logistics Quarterly, 8, 1961, p.175.

(8) Bain, L. J., "Inferences Based on Censored Sampling from the Weibull or Extreme-Value Distribution," Technometrics, 14, 1972, p.693.

(9) Bain, L. J. and Weeks, D. L., "Tolerance Limits for the Generalized Gamma Distribution," American Statistical Association. Journal, 60, 1965, p.1142.

(10) Balinski, M. L. and Gomory, R., "A Primal-Dual Simplex Method," in Recent Advances in Mathematical Programming, McGraw-Hill, New York, 1963.

(11) Bard, Y., "On a Numerical Instability of Davidon-like Methods," Mathematics of Computation, 22, 1968, p.665.

(12) Bard, Y., "Comparison of Gradient Methods for the Solution of Nonlinear Parameter Estimation Problems," Society for Industrial and Applied Mathematics. Journal, 7, 1970, p.157.

(13) Barnes, J. G. P., "An Algorithm for Solving Non-Linear Equations Based on the Secant Method," Computer Journal, 8, 1965, p.66.

(14) Barnett, V. D., "Evaluation of the Maximum-Likelihood Estimator Where the Likelihood Equation has Multiple Roots," Biometrika, 53, 1966, p.151.

(15) Bayes, T., "Essay Towards Solving a Problem in the Doctrine of Chances," Philosophical Transactions of the Royal Society, 53, 1763, p.370.

(16) Beale, E. M. L., "On Minimizing a Convex Function Subject to Linear Inequalities, Royal Statistical Society. Journal, 17, 1955, p.173.

(17) Beale, E. M. L., "On Quadratic Programming," Naval Research Logistics Quarterly, 6, 1959, p.227.

(18) Beltrami, E. J. and McGill, R., "A Class of Variational Problems in Search Theory and the Maximum Principle," Operations Research, 14, 1966, p.267.

(19) Berge, C. Topological Spaces (translated from 1959 French edition) Oliver and Boyd, London, 1963.

(20) Berkson, J., "A Statistically Precise and Relatively Simple Method of Estimating the Bio-Assay with Quantal Response, Based on the Logistic Function," American Statistical Association. Journal, 48, 1953, p.565.

(21) Bernoulli, J., "Essai d'une Nouvelle Theorie de la Manoeuvre des Vasseaux," 1714 (French).

(22) Berrettoni, J. N., "Practical Applications of the Weibull Distribution," Industrial Quality Control, 21, 1964, p.71.

(23) Billman, B. R., Antle, C. L. and Bain, L. J., "Statistical Inference from Censored Weibull Samples," Technometrics, 14, 1971, p.831.

(24) Blackwell, D., "Conditional Expectation and Unbiased Sequential Estimation," Annals of Mathematical Statistics, 18, 1947, p.105.

(25) Bocharov, N. and Fel'dbaum, A. A., "An Automatic Optimizer for the Search of the Smallest of Several Minima," Automation and Remote Control, 23, 1962, p.260.

(26) Booth, G. W. and Peterson, T. I., "Non-Linear Estimation," International Business Machines Corporation Mathematics and Applications Report A & M 3, New York, 1960.

(27) Box, M. J., "A New Method of Constrained Optimization and a Comparison with Other Methods," Computer Journal, 8, 1965, p.42.

(28) Brent, R. P., Algorithms for Minimization Without Derivatives, Prentice-Hall, Englewood Cliffs, New Jersey, 1973.

(29) Brooks, S. H., "A Discussion of Random Methods for Seeking Maxima," Operations Research, 6, 1958, p.249.

(30) Brooks, S. H., "A Comparison of Maximum-Seeking Methods," Operations Research, 7, 1959, p.430.

(31) Brown, G. G. and Rutemiller, H. C., "The Efficiencies of Maximum Likelihood and Minimum Variance Unbiased Estimators of Fraction Defective in the Normal Case," Technometrics, 15, 1973, p.849.

(32) Brown, K. M. and Conte, S., "The Solution of Simultaneous Nonlinear Equations," Association for Computing Machinery. Proceedings of the National Meeting, 1967, p.111.

(33) Brown, K. M. and Dennis, J. E., "On Newton-Like Iteration Functions: General Convergence Theorems and a Specific Algorithm," Numerische Mathematik, 12, 1968, p.186.

(34) Broyden, C. G., "A Class of Methods for Solving Nonlinear Simultaneous Equations," Mathematics of Computation, 19, 1965, p.577.

(35) Broyden, C. G., "Quasi-Newton Methods, and Their Application to Function Minimization," Mathematics of Computation, 21, 1967, p.368.

(36) Broyden, C. G., "A New Method of Solving Nonlinear Simultaneous Equations," Computer Journal, 12, 1969, p.94.

(37) Broyden, C. G., "The Convergence of a Class of Double-Rank Minimization Algorithms, Part I," Institute for Mathematics and Applications. Journal, 1, 1970, p.76.

(38) Broyden, C. G., "The Convergence of a Class of Double-Rank Minimization Algorithms, Part II," Institute for Mathematics and Applications. Journal, 1, 1970, p.222.

(39) Broyden, C. G., "The Convergence of Single-Rank Quasi-Newton Methods," Mathematics of Computation, 24, 1970, p.365.

(40) Buck, R. C. Advanced Calculus, McGraw-Hill, New York, 1965.

(41) Butler, T. and Martin, A. V., "On a Method of Courant for Minimizing Functionals," Journal of Mathematics and Physics, 41, 1962, p.291.

(42) Camp, G. D., "Inequality-Constrained Stationary Value Problems," Journal of the Operations Research Society of America, 3, 1955, p.548.

(43) Cantrel, J. W., "Relation Between the Memory Gradient Method and the Fletcher-Powell Method," Journal of Optimization Theory and Applications, 4, 1969, p.67.

(44) Carroll, C. W., "The Created Response Surface Technique for Optimizing Nonlinear, Restrained, Systems," Operations Research, 9, 1961, p.169.

(45) Cauchy, A. L., "Method Generale pour la Resolution des Systemes d'Equations Simultanees," Comptes Rendus, Academie Science, Paris, 25, 1847, p.536 (French).

(46) Chanda, K. C., "A Note on the Consistency and Maxima of the Roots of the Likelihood Equations," Biometrika, 41, 1954, p.56.

(47) Cheney, E. W. and Goldstein, A. A., "Newton's Method of Convex Programming and Tchebycheff Approximation," Numerische Mathematik, 1, 1959, p.253.

(48) Choi, S. C. and Wette, R., "Maximum Likelihood Estimation of the Parameters of the Gamma Distribution and Their Bias," Technometrics, 11, 1969, p.683.

(49) Clasen, R. J., Graves, G. W. and Lu, J. Y., "Sortie Allocation by a Nonlinear Programming Model for Determining Munitions Mix," Rand Corporation Report R-1411-DDPAE, 1974.

(50) Clough, D. J., "An Asymptotic Extreme Value Sampling Theory for the Estimation of a Global Maximum," Canadian Operations Research Society. Journal, 5, 1969, p.103.

(51) Cochran, W. G., "The Distribution of Quadratic Forms in a Normal System, with Applications to Analysis of Covariance," Cambridge Philosophical Society. Proceedings, 30, 1934, p.178.

(52) Cohen, A. C., "Progressively Censored Samples in Life Testing," Technometrics, 5, 1963, p.327.

(53) Cohen, A. C., "Maximum Likelihood Estimation in the Weibull Distribution Based on Complete and on Censored Samples," Technometrics, 7, 1965, p.579.

(54) Colville, A. R., "A Comparative Study of Nonlinear Programming Codes," International Business Machines New York Scientific Center Technical Report 320-2949, 1968.

(55) Cooper, L., "Heuristic Methods for Location-Allocation," Society for Industrial and Applied Mathematics. Review, 1, 1964, p.37.

(56) Courant, R., "Variational Methods in the Solution of Problems of Equilibrium and Vibrations," American Mathematical Society. Bulletin, 49, 1943, p.1.

(57) Cragg, E. E. and Levy, A. V., "Study of a Supermemory Gradient Method for the Minimization of Functions," Journal of Optimization Theory and Applications, 4, 1969, p.191.

(58) Cramer, H. Mathematical Methods of Statistics Princeton University Press, Princeton, New Jersey, 1946.

(59) Curry, H. D., "The Method of Steepest Descent for Non-Linear Minimization Problems," Quarterly of Applied Mathematics, 2, 1944, p.258.

(60) Daniel, J. W., "Convergence of the Conjugate Gradient Method with Computationally Convenient Modifications," Numerische Mathematik, 10, 1967, p.125.

(61) Daniel, J. W., "The Conjugate Gradient Method for Linear and Nonlinear Operator Equations," Society for Industrial and Applied Mathematics. Journal on Numerical Analysis, 4, 1967, p.10.

(62) Davidenko, D. F., "On a New Method of Numerical Solution of Systems of Nonlinear Equations," Akademiia Nauk SSSR. Doklady, 88, 1953, p.601 (Russian).

(63) Davidon, W. C., "Variable Metric Method for Minimization," Atomic Energy Commission Research and Development Report ANL-5990, 1959.

(64) Davidon, W. C., "Variance Algorithm for Minimization," Computer Journal, 10, 1967, p.406.

(65) Davies, D., "Some Practical Methods for Optimization," in Abadie[1, p.87].

(66) Davies, O. Design and Analysis of Industrial Experiments, Oliver and Boyd, London, 1956.

(67) Davis, H. T. Tables of the Higher Mathematical Functions, Volume 1, Principia Press, Bloomington, Indiana, 1935.

(68) Davis, H. T. Tables of the Higher Mathematical Functions, Volume 2, Principia Press, Bloomington, Indiana, 1935.

(69) Davis, P. J., "Gamma Function and Related Functions," in Handbook of Mathematical Functions edited by M. Abramowitz and I. Stegun, National Bureau of Standards, Washington, D. C., 1964, p.253.

(70) DeMont, R. and White, T., "Analysis of the Combat Empirical Tank Damage Data of the October, 1973 War," MS Thesis in Operations Research, Naval Postgraduate School, Monterey, 1974.

(71) Dennis, J. E., "On Newton-Like Methods," Numerische Mathematik, 11, 1968, p.324.

(72) Dubey, S. D., "On Some Statistical Inferences for Weibull Laws," Naval Research Logistics Quarterly, 13, 1966, p.227.

(73) Dubey, S. D., "Hyper-efficient Estimator of the Location Parameter of the Weibull Laws," Naval Research Logistics Quarterly, 13, 1966, p.253.

(74) Dubey, S. D., "Asymptotic Efficiencies of the Moment Estimators for the Parameters of the Weibull Laws," Naval Research Logistics Quarterly, 13, 1966, p.265.

(75) Dubey, S. D., "Some Percentile Estimators of Weibull Parameters," Technometrics, 9, 1967, p.119.

(76) Dubey, S. D., "On Some Permissible Estimators of the Location Parameter of the Weibull and Certain Other Distributions," Technometrics, 9, 1967, p.293.

(77) Dubey, S. D., "A Compound Weibull Distribution," Naval Research Logistics Quarterly, 15, 1968, p.179.

(78) Faddeeva, V. N. Computational Methods of Linear Algebra, Dover, New York, 1959.

(79) Fiacco, A. V., "A General Regularized Sequential Unconstrained Minimization Technique," Society for Industrial and Applied Mathematics. Journal of Applied Mathematics, 17, 1969, p.1239.

(80) Fiacco, A. V. and Jones, A. P., Generalized Penalty Methods in Topological Spaces," Society for Industrial and Applied Mathematics. Journal of Applied Mathematics, 17, 1969, p.996.

(81) Fiacco, A. V. and McCormick, G. P., "The Sequential Unconstrained Minimization Technique for Nonlinear Programming: A Primal-Dual Method," Management Science, 10, 1964, p.360.

(82) Fiacco, A. V. and McCormick, G. P., "Computational Algorithm for the Sequential Unconstrained Minimization Technique for Nonlinear Programming," Management Science, 10, 1964, p.601.

(83) Fiacco, A. V. and McCormick, G. P. Non-linear Programming: Sequential Unconstrained Minimization Techniques, Wiley, New York, 1968.

(84) Finney, D. J. Statistical Methods in Biological Assay, Griffin, London, 1952.

(85) Fisher, R. A., "The Use of Multiple Measurements in Taxonomic Problems," Annals of Eugenics, 7, 1936, p.179.

(86) Fisher, R. A. Contributions to Mathematical Statistics, Wiley, New York, 1950.

(87) Fletcher, R., "Function Minimization Without Evaluating Derivatives - a Review," Computer Journal, 8, 1965, p.33.

(88) Fletcher, R. Optimization, Academic Press, New York, 1969.

(89) Fletcher, R. and Powell, M. J. D., "A Rapidly Convergent Descent Method for Minimization," Computer Journal, 6, 1963, p.163.

(90) Fletcher, R. and Reeves, C. M., "Function Minimization by Conjugate Gradients," Computer Journal, 7, 1964, p.149.

(91) Forsythe, G. E., "Computing Constrained Minima with Lagrange Multipliers," Society for Industrial and Applied Mathematics. Journal, 3, 1955, p.173.

(92) Forsythe, G. E., "On the Asymptotic Directions of the s-dimensional Optimum Gradient Method," Numerische Mathematik, 11, 1968, p.57.

(93) Frank, M. and Wolfe, P., "An Algorithm for Quadratic Programming," Naval Research Logistics Quarterly, 3, 1956, p.95.

(94) Friedman, P. and Pinder, L. K., "Optimization of a Simulation Model of a Chemical Plant," Industrial and Engineering Chemistry. Process Design and Development, 11, 1972, p.572.

(95) Gauss, K. F., "Theoria Motus Corporum Coelestium," Werke, 7, 1809, p.240 (Latin).

(96) Gauss, K. F., "Disquisitiones Generales Circa Seriem Infinitam," Werke, 3, 1813, p.125 (Latin).

(97) Gill, P. E. and Murray, W., "Quasi-Newton Methods for Unconstrained Optimization," Association for Computing Machinery. Journal, 9, 1972, p.91.

(98) Goldfarb, D., "Extension of Davidon's Variable Metric Method to Maximization Under Linear Inequality and Equality Constraints," Society for Industrial and Applied Mathematics. Journal, 17, 1969, p. .

(99) Goldfarb, D., "A Family of Variable-Metric Methods Derived by Variational Means," Mathematics of Computation, 24, 1970, p.23.

(100) Goldstein, A. A., "Cauchy's Method for Minimization," Numerische Mathematik, 4, 1962, p.146.

(101) Goldstein, A. A., "On Steepest Descent," Society for Industrial and Applied Mathematics. Journal on Control, 3, 1965, p.147.

(102) Goldstein, A. A. and Kripke, B., "Mathematical Programming by Minimizing Differential Functions," Numerische Mathematik, 6, 1962, p.47.

(103) Goldstein, A. A. and Price, J. F., "An Effective Algorithm for Minimization," Numerische Mathematik, 10, 1967, p.184.

(104) Graves, G. W., "Development and Testing of a Non-Linear Programming Algorithm," Aerospace Corporation Report ATR-64(7040)-2, 1964.

(105) Graves, G. W., "A Complete Constructive Algorithm for the General Mixed Linear Programming Problem," Naval Research Logistics Quarterly, 12, 1965, p.1.

(106) Graves, G. W. and Whinston, A. B., "The Application of a Nonlinear Algorithm to a Second-Order Representation of the Problem," Centre d'Etudes de Recherche Operationelle, 11, 1969, p.75.

(107) Greenstadt, J. L., "A Richocheting Gradient Method for Nonlinear Optimization," Society for Industrial and Applied Mathematics. Journal, 14, 1966, p.429.

(108) Greenstadt, J. L., "On the Relative Efficiency of Gradient Methods," Mathematics of Computation, 21, 1967, p.360.

(109) Greenstadt, J. L., "Variations on Variable Metric Methods," Mathematics of Computation, 24, 1970, p.1.

(110) Gumbel, E. Statistics of Extremes, Columbia University Press, New York, 1958.

(111) Hadley, G. Nonlinear and Dynamic Programming, Addison-Wesley, Reading, Massachusetts, 1964.

(112) Hager, H. W., Bain, L. J. and Antle, C. E., "Reliability Estimation for the Generalized Gamma Distribution and Robustness of the Weibull Model," Technometrics, 13, 1971, p.547.

(113) Halperin, M., "Maximum Likelihood Estimation in Truncated Samples," Annals of Mathematical Statistics, 23, 1960, p.55.

(114) Hardy, G. H., Littlewood, J. E. and Polya, G. Inequalities, Cambridge University Press, Cambridge, 1959.

(115) Harter, H. L. and Moore, A. H., "Point and Interval Estimates, Based on m-order Statistics, for the Scale Parameter of a Weibull Population with Known Shape Parameter," Technometrics, 7, 1965, p.405.

(116) Harter, H. L., and Moore, A. H. Maximum Likelihood Estimation of the Parameters of Gamma and Weibull Populations from Complete and from Censored Samples, Technometrics, 7, 1965, p.639.

(117) Harter, H. L. and Moore, A. H., "Local Maximum-Likelihood Estimation of the Parameters of Three-Parameter Lognormal Populations From Complete and Censored Samples," American Statistical Association, Journal, 61, 1966, p.842.

(118) Harter, H. L. and Moore, A. H., "Asymptotic Variances and Covariances of Maximum Likelihood Estimates, From Censored Samples, of the Parameters of Weibull and Gamma Populations," Annals of Mathematical Statistics, 38, 1967, p.557.

(119) Hartley, H. O. and Pfaffenberger, R. C., "Statistical Control of Optimization," in Optimizing Methods in Statistics edited by J. S. Rustagi, Academic Press, New York, 1971, p.281.

(120) Hartman, J. K., "Some Experiments in Global Optimization," Naval Postgraduate School Report NPS55HH72C5A, 1972.

(121) Hatfield, G. B., "A Primal-Dual Method for Minimization with Linear Constraints," Naval Personnel and Training Research Laboratory Technical Bulletin STB73-4, 1973.

(122) Hatfield, G. B. and Graves, G. W., "Optimization of a Reverse Osmosis System Using Nonlinear Programming," Desalinization, 7, 1970, p.147.

(123) Henrici, P. Elements of Numerical Analysis, Wiley, New York, 1964.

(124) Hestenes, M. R. Calculus of Variations and Optimal Control Theory, Wiley, New York, 1966.

(125) Hestenes, M. R., "Multiplier and Gradient Methods," Journal of Optimization Theory and Applications, 4, 1969, p.303.

(126) Hestenes, M. R. and Stiefel, E., "Methods of Conjugate Gradients for Solving Linear Systems," U.S. National Bureau of Standards. Journal of Research, 49, 1952, p.409.

(127) Hildebrand, F. B. Methods of Applied Mathematics, Prentice-Hall, Englewood Cliffs, New Jersey, 1952.

(128) Hogg, R. V. and Craig, A. T. Introduction to Mathematical Statistics, Macmillan, Toronto, Ontario, 1970.

(129) Hooke, R. and Jeeves, T. A., "Direct Search Solution of Numerical and Statistical Problems," Association for Computing Machinery. Journal, 8, 1961, p.212.

(130) Householder, A. S. The Theory of Matrices in Numerical Analysis, Blaisdell, New York, 1964.

(131) Huber, P. J., "Robust Estimation of a Location Parameter," Annals of Mathematical Statistics, 35, 1964, p.73.

(132) Huzurbazar, V. S., "On a Property of Distributions Admitting Sufficient Statistics," Biometrika, 36, 1949, p.71.

(133) Isaacson, E. and Keller, H. B. Analysis of Numerical Methods, Wiley, New York, 1966.

(134) Iverson, K. E. A Programming Language, Wiley, New York, 1962.

(135) John, F., "Extremum Problems with Inequalities as Side Conditions," in Studies and Essays, Courant Anniversary Volume, edited by K. O. Friedrichs, et. al., Wiley, New York, 1948, p.187.

(136) Jones, A., "Spiral - A New Algorithm for Non-Linear Parameter Estimation Using Least Squares," Computer Journal, 13, 1970, p.301.

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

(138) Kale, B. K., "On the Solution of Likelihood Equations by Iteration Processes. The Multiparametric Case," Biometrika, 49, 1962, p.479.

(139) Kao, J. H., "A Graphical Estimation of Mixed Weibull Parameters in Life Testing of Electron Tubes," Technometrics, 1, 1959, p.389.

(140) Kelley, J. E., "The Cutting Plane Method for Solving Convex Programs," Society for Industrial and Applied Mathematics, Journal, 8, 1960, p.703.

(141) Kendall, M. G. and Stuart, A. The Advanced Theory of Statistics, Volume 2, Inference and Relationship, Griffin, London, 1967.

(142) Koopman, B. O., "On Distributions Admitting a Sufficient Statistic," American Mathematical Society. Transactions, 39, 1936, p.399.

(143) Kortanek, K. O. and Evans, J. P., "Psuedo-Concave Programming and Lagrangian Regularity," Operations Research, 15, 1967, p.882.

(144) Kowalik, J. S. and Osborne, M. R. Methods for Unconstrained Optimization, Elsevier, New York, 1968.

(145) Kowalik, J. S., Osborne, M. R. and Ryan, D. M., "A New Method for Constrained Optimization Problems," Operations Research, 17, 1969, p.973.

(146) Kuhn, H. W. and Tucker, A. W., "Nonlinear Programming," Berkeley Symposium on Mathematical Statistics and Probability. Proceedings, University of California Press, 1951, p.481.

(147) Lagrange, C. J., "Essai d'une Nouvelle Methode pour Determiner les Maxima et les Minima," 1760 (French).

(148) Lancaster, P., "Error Analysis of the Newton-Raphson Method," Numerische Mathematik, 9, 1966, p.55.

(149) Le Cam, L., "On the Assumptions Used to Prove Asymptotic Normality of Maximum Likelihood Estimates," Annals of Mathematical Statistics, 41, 1970, p.802.

(150) Legendre, A. M., "Nouvelles Methodes pour la Determination des Orbites de Cometes," Paris, 1805 (French).

(151) Lehman, E. H., "Shapes, Moments and Estimation of the Weibull Distribution," Institute of Electronic and Electrical Engineers. Transactions on Reliability, R-12, 1963, p.32.

(152) Lemke, C. E., "A Method of Solution for Quadratic Programs," Management Science, 8, 1962, p.442.

(153) Levenberg, K., "A Method for the Solution of Certain Non-Linear Problems in Least Squares," Quarterly of Applied Mathematics, 2, 1944, p.164.

(154) Liau, T. L., Hartley, H. O. and Sielken, R. L., "Confidence Regions for Global Optima in Nonlinear Programming," Texas A&M University Project Themis Technical Report 43, 1973.

(155) Lill, S. A., "A Modified Davidon Method for Finding the Minimum of a Function Using Difference Approximations for Derivatives, Algorithm 46," Computer Journal, 13, 1970, p.111.

(156) Lloyd, D. K. and Lipow, M. Reliability: Management, Methods, and Mathematics, Prentice-Hall, Englewood Cliffs, New Jersey, 1962.

(157) McCool, J., "Inferences on Weibull Percentiles and Shape Parameter from Maximum Likelihood Estimates," Institute of Electronic and Electrical Engineers. Transactions on Reliability, R-19, 1970, p.2.

(158) Mangasarian, O. L. Nonlinear Programming, McGraw-Hill, New York, 1969.

(159) Mangasarian, O. and Fromovitz, S., "The Fritz John Necessary Conditions in the Presence of Equality and Inequality Constraints," Journal of Mathematical Analysis and Applications, 17, 1967, p.37.

(160) Mann, N. R., "Point and Interval Estimation Procedures for the Two-Parameter Weibull and Extreme-Value Distributions," Technometrics, 10, 1968, p.231.

(161) Mann, N., Schafer, R. and Singpurwalla, N. Methods for Statistical Analysis of Reliability and Life Data, Wiley, New York, 1974.

(162) Mantel, N. and Myers, M., "Problems of Convergence of Maximum Likelihood Iterative Procedures in Multiparameter Situations," American Statistical Association. Journal, 66, 1971, p.484.

(163) Markowitz, H., "The Optimization of a Quadratic Function Subject to Linear Constraints," Naval Research Logistics Quarterly, 3, 1956, p.111.

(164) Marquardt, D. W., "An Algorithm for Least Squares Estimation of Nonlinear Parameters," Society for Industrial and Applied Mathematics. Journal, 11, 1963, p.431.

(165) Matthews, A. and Davies, D., "A Comparison of Modified Newton Methods for Unconstrained Optimization," Computer Journal, 14, 1971, p.293.

(166) Menon, M. V., "Estimation of the Shape and Scale Parameters of the Weibull Distribution," Technometrics, 5, 1963, p.175.

(167) Michelini, C., "Convergence Pattern of the Scoring Method in Estimating Parameters of a Log-Normal Function," American Statistical Association. Journal, 67, 1972, p.319.

(168) Miele, A. and Cantrell, J. W., "Study of a Memory Gradient Method for the Minimization of Functions," Journal of Optimization Theory and Applications, 3, 1969, p.459.

(169) Mood, A. and Graybill, F. Introduction to the Theory of Statistics, McGraw-Hill, New York, 1963.

(170) Myers, G. E., "Properties of the Conjugate Gradient and Davidon Methods," Journal of Optimization Theory and Applications, 2, 1968, p.209.

(171) Nelder, J. A. and Mead, R., "A Simplex Method for Function Minimization," Computer Journal, 7, 1965, p.308.

(172) Neyman, J. A Selection of Early Statistical Papers of J. Neyman, Cambridge University Press, Cambridge, England, 1967.

(173) Neyman, J. and Pearson, E. S., "On the Use and Interpretation of Certain Test Criteria for Purposes of Statistical Inference," Biometrica, 20A, 1928, p.175, 263.

(174) Neyman, J. and Pearson, E. S. Joint Statistical Papers of J. Neyman and E. S. Pearson, Cambridge University Press, Cambridge, England, 1967.

(175) Nilson, N. Learning Machines, McGraw-Hill, New York, 1965.

(176) Nilson, N. Problem Solving Methods in Artificial Intelligence, McGraw-Hill, New York, 1971.

(177) Noh, J. C., "A Two Phase Complex Method for Non-Linear Process Optimization," presented at ORSA/TIMS 45th National Meeting, Boston, 1974.

(178) Oren, S. and Luenberger, D., "Self-Scaling Variable Metric (SSVM) Algorithms, Part I: Criteria and Sufficient Conditions for Scaling a Class of Algorithms," Management Science, 20, 1974, p.845.

(179) Oren, S. and Luenberger, D., "Self-Scaling Variable Metric (SSVM) Algorithms, Part II: Implementation and Experiments," Management Science, 20, 1974, p.863.

(180) Page, J., "Automated Design of Feedback Control Systems," PHD Dissertation in Engineering, University of California, Los Angeles, 1970.

(181) Parkinson, J. M. and Hutchinson, D., "An Investigation into the Efficiency of Variants of the Simplex Method," in Numerical Methods for Nonlinear Optimization, edited by F. A. Lootsma, Academic Press, New York, 1972.

(182) Pearson, J. D., "Variable Metric Methods of Minimization," Computer Journal, 12, 1969, p.171.

(183) Pearson, K. Karl Pearson's Early Statistical Papers, Cambridge University Press, Cambridge, England, 1948.

(184) Peto, R. and Lee, P., "Weibull Distributions for Continuous-Carcinogenesis Experiments," Biometrics, 29, 1973, p.457.

(185) Pike, M., "A Suggested Method of Analysis of a Certain Class of Experiments in Carcinogenesis," Biometrics, 29, 1966, p.142.

(186) Pitman, E. J. G., "Sufficient Statistics and Intrinsic Accuracy," Cambridge Philosophical Society. Proceedings, 32, 1936, p.567.

(187) Polfeldt, T., "The Order of the Minimum Variance in a Non-Regular Case," Annals of Mathematical Statistics, 1, 1970, p.667.

(188) Polfeldt, T., "Minimum Variance Order When Estimating the Location of an Irregularity in the Density," Annals of Mathematical Statistics, 41, 1970, p.673.

(189) Pomentale, T., "A New Method for Solving Conditioned Maxima Problems," Journal of Mathematical Analysis and Applications, 10, 1965, p.216.

(190) Powell, M. J. D. "An Efficient Method for Finding the Minimum of a Function of Several Variables Without Calculating Derivatives," Computer Journal, 7, 1964, p.155.

(191) Powell, M. J. D., "Rank One Methods for Unconstrained Minimization," Atomic Energy Research Establishment Report T.P. 372, 1969 (also in Abadie[1], p.139).

(192) Powell, M. J. D., "A Survey of Numerical Methods for Unconstrained Optimization," Society for Industrial and Applied Mathematics. Review, 12, 1970, p.79.

(193) Ralston, A. First Course in Numerical Analysis, McGraw-Hill, New York, 1965.

(194) Ramsay, J. O., "A Family of Gradient Methods for Optimization," Computer Journal, 13, 1970, p.413.

(195) Rao, C. R., "Information and Accuracy Attainable in the Estimation of Statistical Parameters," Calcutta Mathematical Society. Bulletin, 37, 1945, p.81.

(196) Rao, C. R., "Theory of the Method of Estimation by Minimum Chi-Square, International Statistical Institute. Bulletin, 35, 1957, p.25.

(197) Rao, C. R. Linear Statistical Inference and Its Application to Biometric Research, Wiley, New York, 1965.

(198) Ravenis, J. V. J., "Estimating Weibull-Distribution Parameters," Electro-Technology, 73, 1964, p.46.

(199) Richardson, J. A. and Kuester, J. L., "The Complex Method for Constrained Optimization - Algorithm 454," Association for Computing Machinery. Communications, 16, 1973, p.487.

(200) Ringer, L. J. and Sprinkle, E. E., "Estimation of the Parameters of the Weibull Distribution from Multicensored Samples," Institute of Electronic and Electrical Engineers. Transactions on Reliability, R-21, 1972, p.46.

(201) Rockette, H., "Statistical Inference for the Three Parameter Weibull Distribution," PHD Dissertation in Statistics, Pennsylvania State University, 1972.

(202) Rockette, H., Antle, C. and Klimko, L., "Maximum Likelihood Estimation with the Weibull Model," American Statistical Association. Journal (to appear).

(203) Rosen, J. B., "The Gradient Projection Method for Nonlinear Programming, Part I. Linear Constraints," Society for Industrial and Applied Mathematics. Journal, 8, 1960, p.181.

(204) Rosen, J. B., "The Gradient Projection Method for Nonlinear Programming, Part II. Nonlinear Constraints," Society for Industrial and Applied Mathematics. Journal, 9, 1961, p.514.

(205) Rosen, J. B. and Suzuki, S., "Construction of Nonlinear Programming Test Problems," Association for Computing Machinery. Communications, 8, 1965, p.113.

(206) Rosenbrock, H., "An Automatic Method for Finding the Greatest and Least Value of a Function," Computer Journal, 3, 1960, p.175.

(207) Ross, G. J. S., "The Efficient Use of Function Minimization in Non-linear Maximum-likelihood Estimation," Applied Statistics, 19, 1970, p.205.

(208) Scheefer, L., "Uber die Bedeutung der Begriffe Maximum und Minimum in der Variationsrechnung," 1886 (German).

(209) Schwarz, H. R., Rutishauser, H. and Stiefel, E. Numerical Analysis of Symmetric Matrices, Prentice-Hall, Englewood Cliffs, New Jersey, 1973.

(210) Shah, B. V., Buehler, R. J. and Kempthorne, O., "The Method of Parallel Tangents (Partan) for Finding an Optimum," Office of Naval Research Report, NR-042-207, 2, 1961.

(211) Shere, K. D., "Remark on Algorithm 454, The Complex Method for Constrained Optimization," Association for Computing Machinery. Communications, 17, 1974, p.471.

(212) Smith, E. B. and Shanno, D. F., "An Improved Marquardt Procedure for Nonlinear Regressions," Technometrics, 13, 1971, p.63.

(213) Smith, H. and Dubey, S. D., "Some Reliability Problems in the Chemical Industry," Industrial Quality Control, 21, 1964, p.64.

(214) Solberg, E., "Labor Supply and Labor Force Participation Decisions of the AFDC Population-at-Risk," PHD Dissertation in Economics, Claremont Graduate School, 1974.

(215) Spang, H. A., "A Review of Minimization Techniques for Nonlinear Functions," Society for Industrial and Applied Mathematics. Review, _, 1962, p.343.

(216) Spendley, W., Hext, G. R., and Himsworth, F. R., "Sequential Application of Simplex Designs in Optimisation and Evolutionary Operation," Technometrics, 4, 1962, p.441.

(217) Sprott, D. and Kalbfleisch, J., "Examples of Likelihoods and Comparison with Point Estimations and Large Sample Approximations," American Statistical Association, Journal, 64, 1969, p.468.

(218) Stewart, G. W., "A Modification of Davidon's Minimization Method to Accept Difference Approximations," Association for Computing Machinery, Journal, 14, 1967, p.72.

(219) Stong, R., "A Note on the Sequential Unconstrained Minimization Technique for Non-Linear Programming," Management Science, 12, 1965, p.142.

(220) Theil, H. and van de Panne, C., "Quadratic Programming as an Extension of Classical Quadratic Maximization," Management Science, 7, 1960, p.1.

(221) Thoman, D. R., Bain, L. J. and Antle, C. E., "Inferences on the Parameters of the Weibull Distribution," Technometrics, 11, 1969, p.445.

(222) van de Panne, C. and Whinston, A., "The Simplex and Dual Method for Quadratic Programming," Operational Research Quarterly, 15, 1964, p.355.

(223) Vandaele, W. H. and Chowdhury, S. R., "A Revised Method of Scoring," Statistica Neerlandica, 25, 1971, p.101.

(224) Wald, A., "Tests of Statistical Hypotheses Concerning Several Parameters when the Number of Observations is Large," American Mathematical Society, Transactions, 54, 1943, p.426.

(225) Wald, A. "Note on the Consistency of the Maximum Likelihood Estimate," Annals of Mathematical Statistics, 14, 1949, p.595.

(226) Walker, S. H. and Duncan, D. B., "Estimation of the Probability of an Event as a Function of Several Independent Variables," Biometrika, 54, 1967, p.167.

(227) Weibull, W., "A Statistical Theory of Strength of Materials," Ingeniorsvetenskapsakademien Handlingar, N. 151, 1939.

(228) Weibull, W., "The Phenomenum of Rupture in Solids," Ingeniorsvetenskapsakademien Handlingar, N. 153, 1939.

(229) Weibull, W. "A Statistical Distribution of Wide Applicability," Journal of Applied Mechanics, 18, 1951, p.293.

(230) Wilde, D. J., "Differential Calculus in Non-Linear Programming," Operations Research, 10, 1962, p.764.

(231) Wilde, D. J., "Jacobians in Constrained Non-Linear Optimization," Operations Research, 13, 1965, p.848.

(232) Wilde, D. J. and Beightler, C. Foundations of Optimization, Prentice-Hall, Englewood Cliffs, New Jersey, 1967.

(233) Wingo, D. R., "Maximum Likelihood Estimation of the Parameters of the Weibull Distribution by Modified Quasilinearization," Institute of Electronic and Electrical Engineers. Transactions on Reliability, R-21, 1972, p.89.

(234) Wingo, D. R., "Solution of the Three-Parameter Weibull Equations by Constrained Modified Quasilinearization (Progressively Censored Samples)," Institute of Electronic and Electrical Engineers. Transactions on Reliability, R-22, 1973, p.96.

(235) Wolfe, P., "The Secant Method for Solving Nonlinear Equations," Association for Computing Machinery. Communications, 2, 1959, p.12.

(236) Wolfe, P., "The Simplex Method for Quadratic Programming," Econometrica, 27, 1959, p.382.

(237) Wolfe, P., "Accelerating the Cutting Plane Method for Non-linear Programming," Society for Industrial and Applied Mathematics. Journal, 9, 1961, p.481.

(238) Zakharov, V. V., "A Random Search Method," Engineering Cybernetics, 2, 1969, p.26.

(239) Zangwill, W. I., "Minimizing a Function Without Calculating Derivatives," Computer Journal, 10, 1967, p.293.

(240) Zangwill, W. I., "The Convex Simplex Method," Management Science, 14, 1967, p.221

(241) Zangwill, W. I., "Non-Linear Programming Via Penalty Functions," Management Science, 5, 1967, p.344.

(242) Zangwill, W. I., "A Decomposable Nonlinear Programming Approach, " Operations Research, 15, 1967, p.1068.

(243) Zangwill, W. I. Nonlinear Programming: A Unified Approach, Prentice-Hall, Englewood Cliffs, New Jersey, 1969.

(244) Zoutendijk, G. Methods of Feasible Directions, Elsevier, Amsterdam, 1960.