INTRODUCTION TO OPTIMIZATION METHODS AND THEIR APPLICATION IN STATISTICS

B. S. Everitt

Introduction to Optimization Methods and their Application in Statistics

B. S. Everitt BSc MSc

Reader in Statistics in Behavioural Science and
Head of Biometrics Unit, Institute of Psychiatry

London New
CHAPMAN AND HALL

First published in 1987 by Chapman and Hall Ltd
11 New Fetter Lane, London EC4P 4EE
Published in the USA by Chapman and Hall
29 West 35th Street, New York NY 10001

© 1987 B. S. Everitt

Printed in Great Britain by St Edmundsbury Press Ltd
Bury St Edmunds, Suffolk

ISBN 0 412 272105 (HB)
ISBN 0 412 296802 (PB)

This title is available in both hardbound and paperback editions. The paperback edition is sold subject to the condition that it shall not, by way of trade or otherwise, be lent, resold, hired out, or otherwise circulated without the publisher's prior consent in any form of binding or cover other than that in which it is published and without a similar condition including this condition being imposed on the subsequent purchaser.

All rights reserved. No part of this book may be reprinted, or reproduced or utilized in any form or by any electronic, mechanical or other means, now known or hereafter invented, including photocopying and recording, or in any information storage and retrieval system, without permission in writing from the publisher

British Library Cataloguing in Publication Data

Everitt, Brian
Introduction to optimization methods and their applications in statistics.
1. Mathematical statistics 2. Mathematical optimization
I. Title
519.5 QA276.A1

ISBN 0-412-27210-5
ISBN 0-412-29680-2 Pbk

Library of Congress Cataloging in Publication Data

Everitt, Brian.
Introduction to optimization methods and their application in statistics.

Bibliography: p.
Includes index.
1. Mathematical statistics. 2. Mathematical optimization I. Title.
QA276.E92 1987 519.5 87-11693
ISBN 0-412-27210-5
ISBN 0-412-29680-2 (pbk.)

Preface

Optimization techniques are used to find the values of a set of parameters which maximize or minimize some objective function of interest. Such methods have become of great importance in statistics for estimation, model fitting, etc. This text attempts to give a brief introduction to optimization methods and their use in several important areas of statistics. It does not pretend to provide either a complete treatment of optimization techniques or a comprehensive review of their application in statistics; such a review would, of course, require a volume several orders of magnitude larger than this since almost every issue of every statistics journal contains one or other paper which involves the application of an optimization method.

It is hoped that the text will be useful to students on applied statistics courses and to researchers needing to use optimization techniques in a statistical context.

Lastly, my thanks are due to Bertha Lakey for typing the manuscript.

B. S. Everitt
August 1986

Contents

	Preface	vii
1	**An introduction to optimization methods**	1
1.1	Introduction	1
1.2	The optimization problem	2
1.3	Some simple examples	5
1.4	Minimization procedures	7
1.5	Constrained minimization	9
1.6	Summary	10
2	**Direct search methods**	11
2.1	Introduction	11
2.2	Univariate search methods	11
2.3	Multiparameter search methods	16
2.4	Summary	20
3	**Gradient methods**	21
3.1	Introduction	21
3.2	The method of steepest descent	21
3.3	The Newton–Raphson method	23
3.4	The Davidon–Fletcher–Powell method	24
3.5	The Fletcher–Reeves method	25
3.6	Summary	27
4	**Some examples of the application of optimization techniques to statistical problems**	28
4.1	Introduction	28
4.2	Maximum likelihood estimation	28
4.3	Maximum likelihood estimation for incomplete data	35
4.4	Summary	41

此为试读,需要完整PDF请访问: www.ertongbook.com

5 Optimization in regression problems 42
5.1 Introduction 42
5.2 Regression 42
5.3 Non-linear regression 43
5.4 Log-linear and linear logistic models 48
5.5 The generalized linear model 56
5.6 Summary 58

6 Optimization in multivariate analysis 59
6.1 Introduction 59
6.2 Maximum likelihood factor analysis 59
6.3 Cluster analysis 64
6.4 Multidimensional scaling 71
6.5 Summary 77

Appendix: exercises 80

References 83

Index 87

1

An introduction to optimization methods

1.1 INTRODUCTION

A problem considered in all basic statistics courses is that of finding estimates of the two parameters in a simple linear regression model relating a dependent variable, y, to an explanatory variable, x. The model is usually formulated as

$$y_i = \alpha + \beta x_i + \epsilon_i \tag{1.1}$$

where $x_i, y_i, i = 1, \ldots, n$ are the values of the explanatory and dependent variable for a sample of observations considered to arise from the model, and the $\epsilon_i, i = 1, \ldots, n$ are 'error' or residual terms with zero expected values, accounting for how much an observation, y_i, differs from its predicted value, $\alpha + \beta x_i$.

The problem of finding estimates of the parameters, α and β, of the regression model in (1.1) may be approached in several ways. Perhaps the most common is to seek some goodness-of-fit criterion which measures, in some sense, how closely the model agrees with the observed data, and then choose values for the two parameters which *minimize* the chosen measure of fit. An obvious goodness-of-fit criterion for the simple linear regression model is the sum-of-squares of the error terms in (1.1), that is

$$S = \sum_{i=1}^{n} \epsilon_i^2 \tag{1.2}$$

Clearly S does measure how well the observed values of the dependent variable fit those predicted by the model, with smaller values of S indicating a better fit. Consequently choosing as estimates of α and β those values which minimize S is an intuitively reasonable procedure and is, of course, nothing less than the well-known *least squares* estimation technique.

Another commonly occurring estimation problem in statistics arises when we wish to estimate the parameter or parameters of a probability density function given a random sample taken from the density function. For

example, we may have a sample of n values, $x_1, \ldots, x_n$, from an exponential density function of the form

$$f(x) = \lambda e^{-\lambda x} \qquad x > 0 \tag{1.3}$$

and we wish to estimate λ. A very useful estimation procedure in this situation is to form the joint probability density function of the observations, that is

$$\mathscr{L}(x_1, \ldots, x_n; \lambda) = \prod_{i=1}^{n} \lambda e^{-\lambda x_i} \tag{1.4}$$

and choose as the estimate of λ the value which maximizes $\mathscr{L}$, which is generally referred to as the *likelihood function*. This procedure will also be well known to most readers as *maximum likelihood estimation*.

Both the estimation problems described above can be formulated in terms of *optimizing* some numerical function with respect to a number of parameters, and many other statistical problems may be formulated in a similar manner. It is methods for performing such optimizations and their application in statistics which are the main concern of this text.

1.2 THE OPTIMIZATION PROBLEM

In its most general form the problem with which we will be concerned involves finding the optimum value (maximum or minimum) of a function $f(\theta_1, \ldots, \theta_m)$ of m parameters, $\theta_1, \ldots, \theta_m$. We should note at this stage that from a mathematical point of view there is little point in considering both maximization and minimization since maximizing f is equivalent to minimizing $-f$; consequently the discussion in the remainder of the text will normally be confined to minimization. The values taken by the parameters may in some situations be *constrained* and in others *unconstrained*. For example, in the linear regression model of the previous section, the parameters α and β may both take any real value; in other words, they are unconstrained. The parameter of the exponential distribution in (1.3) is, however, constrained to take only positive values. Some comments about the constrained optimization problem will be made in Section 1.5.

Many of the concepts we shall need in discussing optimization methods can be introduced via the case of a function of a single parameter, and Fig. 1.1 shows a graphical representation of such a function. This graph shows that the function has two minima, one at θ_0 and one at θ_1, a maximum at θ_2, and a point of inflexion at θ_3. The minimum at θ_0 is known as a *local minimum* since the value of $f(\theta_0)$ is lower than $f(\theta)$ for values of θ in the neighbourhood of θ_0; the minimum at θ_1 is known as a *global minimum* since $f(\theta_1)$ is lower than $f(\theta)$ for *all* values of θ. As we shall see later, a major problem in complex minimization problems is to decide whether we have found a local or global minimum.

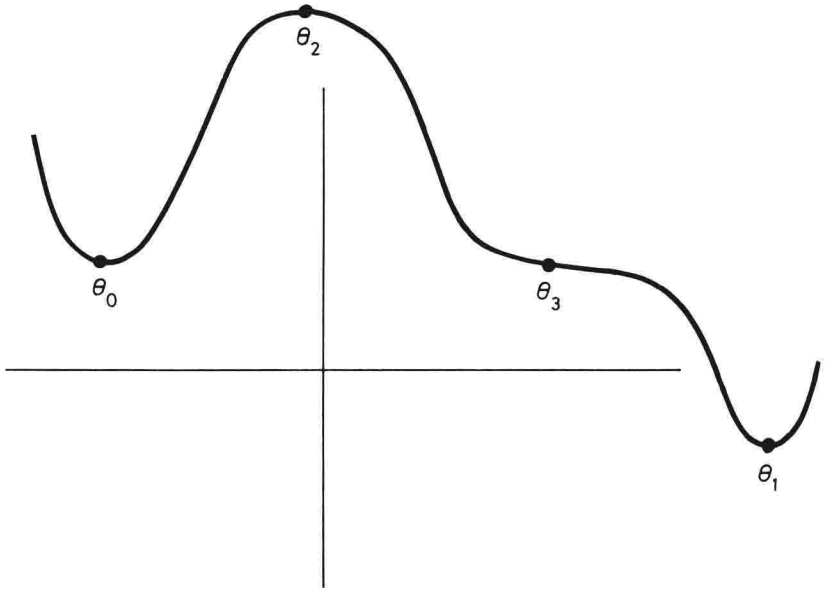

Figure 1.1 Function of a single parameter showing a maximum (θ_2), a local minimum (θ_0), global minimum (θ_1) and point of inflexion (θ_3).

The classical approach to the problem of finding the values θ_0 and θ_1 is to note that at both θ_0 and θ_1 the gradient of $f(\theta)$ is zero, so that θ_0 and θ_1 will be solutions of the equation

$$\frac{df}{d\theta} = 0 \tag{1.5}$$

As we can see from Fig. 1.1 the value θ_2, at which there is a local maximum, and θ_3, at which there is a horizontal point of inflexion, also satisfy this equation; consequently satisfying equation (1.5) is a *necessary* but not a *sufficient* condition for a point to be a minimum. However, examining again Fig. 1.1, we see that at θ_0 and θ_1 the gradient changes sign from negative to positive, at θ_2 the change is from positive to negative, and at θ_3 the gradient does not change sign. So at a minimum the gradient is an increasing function; the rate of change of the gradient is measured by the second derivative so for a minimum we require

$$\frac{d^2f}{d\theta^2} > 0 \tag{1.6}$$

when evaluated at the suspected minimum point.

These ideas may be extended to the minimization of a function of several variables, $f(\theta_1, \ldots, \theta_m)$, so that a necessary condition for a minimum is that

$$\frac{\partial f}{\partial \theta_1} = \frac{\partial f}{\partial \theta_2} = \ldots = \frac{\partial f}{\partial \theta_m} = 0 \tag{1.7}$$

Solutions to these equations may also represent maxima or saddle points, and these various possibilities are illustrated for a function of two variables by the contour diagram shown in Fig. 1.2.

On this diagram, P_1, P_2, P_3 and P_4 are points at which equations (1.7) are satisfied; the corresponding values of f are 0.0, 2.5, 6.5 and 3.5. P_1 is the *global minimum*, that is the required overall minimum of the function. P_2 is a *local minimum*, that is $f(P_2)$ is less than f for all points in the immediate neighbourhood of P_2 but $f(P_2) > f(P_1)$. P_3 is a *local maximum* of f, and P_4 is a *saddle point*; along the direction AB it corresponds to a maximum of f, while along CD it corresponds to a minimum.

The *sufficient* condition for a solution of (1.7) to be a minimum, corresponding to the requirement given in (1.6) for the single-parameter case, is that the matrix **H** with elements h_{ij} given by

$$h_{ij} = \frac{\partial^2 f}{\partial \theta_i \partial \theta_j} \tag{1.8}$$

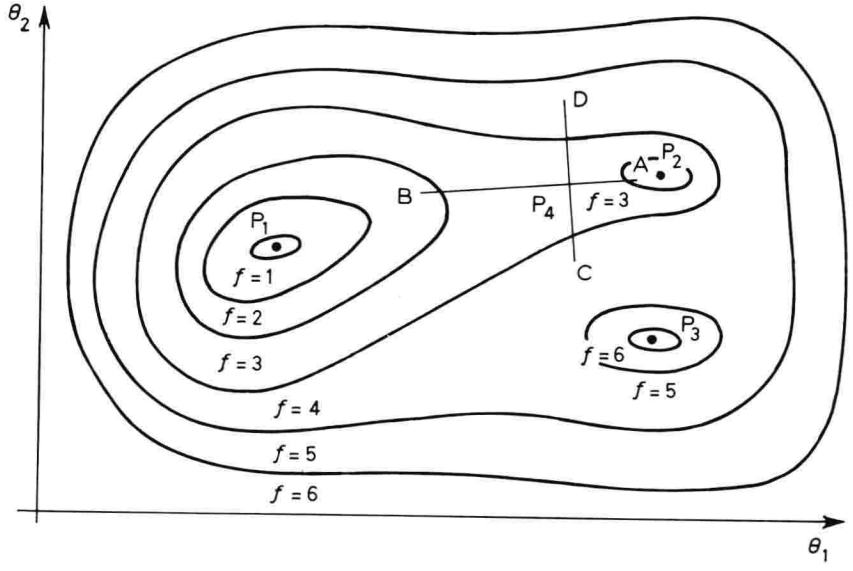

Figure 1.2 Contour diagram of a function of two parameters having a local maximum, a local minimum and a saddle point.

be positive definite when evaluated at the point being considered. **H** is known as the *Hessian* matrix; it is symmetric and of order $m \times m$.

1.3 SOME SIMPLE EXAMPLES

Let us return to the two examples described in Section 1.1 to illustrate a number of the points made in Section 1.2. Consider first the problem of the maximum likelihood estimation of the parameter of an exponential density function. The likelihood function for a sample of n values is given by

$$\mathscr{L}(x_i, \ldots, x_n; \lambda) = \prod_{i=1}^{n} \lambda e^{-\lambda x_i} \tag{1.9}$$

We wish to choose that value of λ which maximizes $\mathscr{L}$ or, equivalently minimizes $-\mathscr{L}$. As with most maximum likelihood problems a simplification is achieved if we consider not $\mathscr{L}$ but the log-likelihood function, L, given by

$$L = \log_e(\mathscr{L}) \tag{1.10}$$

$\mathscr{L}$ and L clearly have their maxima at the same points but in practice L is usually far more convenient to deal with. For the exponential density

$$L = n \log_e \lambda - \lambda \sum_{i=1}^{n} x_i \tag{1.11}$$

Consequently the function we require to minimize is

$$F = \lambda \Sigma x_i - n \log_e \lambda \tag{1.12}$$

Differentiating with respect to λ gives

$$\frac{dF}{d\lambda} = \Sigma x_i - \frac{n}{\lambda} \tag{1.13}$$

Setting $dF/d\lambda$ to zero leads to the following estimator for λ:

$$\hat{\lambda} = n / \sum_{i=1}^{n} x_i \tag{1.14}$$

that is the reciprocal of the sample mean. Clearly this corresponds to a minimum of F since

$$\frac{d^2F}{d\lambda^2} = \frac{n}{\lambda^2} \tag{1.15}$$

which is always positive.

Now let us consider the least squares estimation of the two parameters in

此为试读,需要完整PDF请访问: www.ertongbook.com

the simple linear regression model. This involves minimization of the goodness-of-fit criterion specified in (1.2), which may be rewritten as follows:

$$S = \sum_{i=1}^{n} (y_i - \alpha - \beta x_i)^2 \tag{1.16}$$

so that the equations given by (1.7) take the form

$$\frac{\partial S}{\partial \alpha} = -2 \sum_{i=1}^{n} (y_i - \alpha - \beta x_i) = 0, \tag{1.17}$$

$$\frac{\partial S}{\partial \beta} = -2 \sum_{i=1}^{n} x_i (y_i - \alpha - \beta x_i) = 0 \tag{1.18}$$

Solving these two equations leads to the following well-known estimators for α and β:

$$\hat{\alpha} = \bar{y} - \hat{\beta}\bar{x}, \tag{1.19}$$

$$\hat{\beta} = \frac{Cxy}{Cxx}, \tag{1.20}$$

where

$$Cxy = \sum_{i=1}^{n} x_i y_i - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i / n, \tag{1.21}$$

$$Cxx = \sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i \right)^2 / n, \tag{1.22}$$

The Hessian matrix for this problem is given by

$$\mathbf{H} = \begin{bmatrix} \dfrac{\partial^2 S}{\partial \alpha^2} & \dfrac{\partial^2 S}{\partial \alpha \, \partial \beta} \\ \dfrac{\partial^2 S}{\partial \beta \, \partial \alpha} & \dfrac{\partial^2 S}{\partial \beta^2} \end{bmatrix}, \tag{1.23}$$

$$= \begin{bmatrix} 2n & 2 \sum_{i=1}^{n} x_i \\ 2 \sum_{i=1}^{n} x_i & 2 \sum_{i=1}^{n} x_i^2 \end{bmatrix} \tag{1.24}$$

It is easy to show that **H** is positive definite and consequently that (1.19) and (1.20) correspond to a minimum of the function S.

In both these cases the solutions to the equations specifying the minimum (equations (1.5) and (1.7)) could be solved directly to give estimators for the parameters which were simple functions of the observations. In many situations, however, these equations cannot be solved directly, and other approaches must be adopted to the minimization problem. Some general characteristic of the type of procedure necessary are discussed in the following section; detailed descriptions of specific techniques will be left until Chapters 2 and 3.

1.4 MINIMIZATION PROCEDURES

The minimization techniques to be discussed in the next two chapters all have certain features in common. The most obvious is that they are *iterative* and proceed by generating a sequence of solutions each of which represents an improved approximation to the parameter values at the minimum of f in the sense that

$$f(\boldsymbol{\theta}_{i+1}) \leq f(\boldsymbol{\theta}_i) \tag{1.25}$$

where $\boldsymbol{\theta}_{i+1}$ and $\boldsymbol{\theta}_i$ are vectors containing the values of the m parameters at iterations $i+1$ and i. Such procedures require an initial set of parameter values, $\boldsymbol{\theta}_0$, generally supplied by the investigator, from which successive approximations arise by means of an equation of the form

$$\boldsymbol{\theta}_{i+1} = \boldsymbol{\theta}_i + h_i \mathbf{d}_i. \tag{1.26}$$

In this equation $\mathbf{d}_i$ is an m-dimensional vector specifying the *direction* to be taken in moving from $\boldsymbol{\theta}_i$ to $\boldsymbol{\theta}_{i+1}$ and h_i is a scalar specifying the *distance* to be moved along this direction.

The choice of a suitable direction and distance (often referred to as the *step size*) to ensure that (1.25) is satisfied may be made in a number of ways; it may rely solely on values of the function plus information gained from earlier iterations, or on values of the partial derivatives of f with respect to the parameters. Techniques adopting the first approach are generally known as *direct search methods* and are discussed in Chapter 2. The second type of approach, *gradient methods* are the subject of Chapter 3.

A problem common to all the techniques to be discussed in the next two chapters is how to decide when the iterative procedure has reached the required minimum. In general such decisions are taken on the basis of the sequences $\{\boldsymbol{\theta}_i\}$ and $\{f(\boldsymbol{\theta}_i)\}$, and possible convergence criteria are

$$|f(\boldsymbol{\theta}_{i+1}) - f(\boldsymbol{\theta}_i)| < \epsilon \tag{1.27}$$

and/or

$$||\boldsymbol{\theta}_{i+1} - \boldsymbol{\theta}_i|| < \epsilon' \tag{1.28}$$

for prescribed values of ϵ and ϵ'. Although such criteria are commonly used and are, in many situations, satisfactory, they can in some circumstances cause the iterative procedure to be terminated prematurely. For example, Fig. 1.3 illustrates a case where terminating the iterations on the basis of the fractional changes in $f(\boldsymbol{\theta})$ being less than some small number, causes the procedure to finish on a flat plateau. Figure 1.4 illustrates a case where the use of (1.28) causes premature termination on a very steep slope.

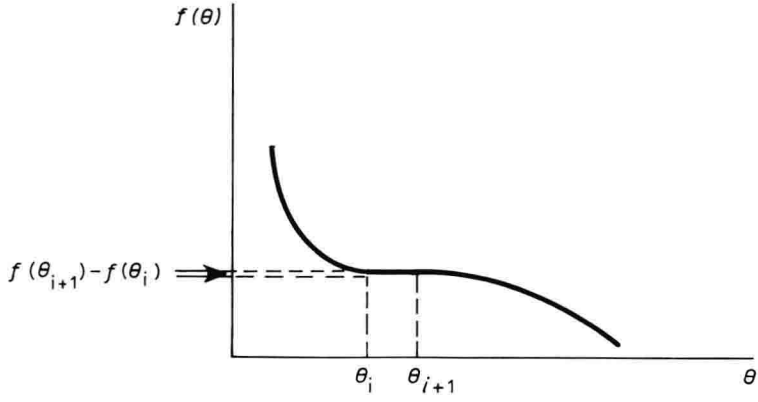

Figure 1.3 Premature termination on a flat plateau.

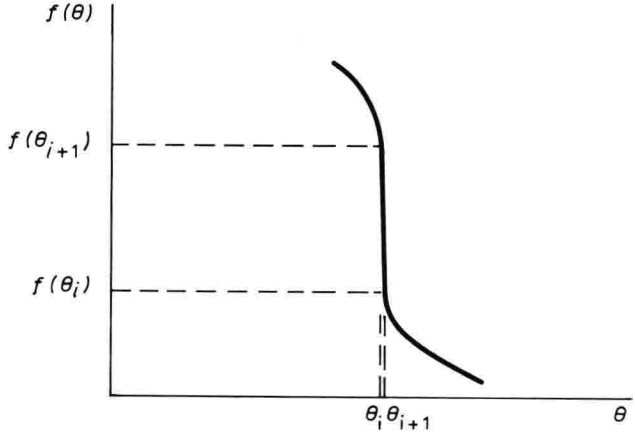

Figure 1.4 Premature termination on a steep slope.

In order to guard against such possibilities a more stringent convergence criterion might be used in which (1.27) or (1.28) were required to hold for each of a number of consecutive iterations.

Termination criteria such as (1.27) and (1.28) are strongly dependent on the scaling of both the objective function, f, and the parameters, $\theta_1, \ldots, \theta_m$. For example, if $\epsilon = 10^{-3}$ and f is always in the interval $(10^{-7}, 10^{-5})$, then it is likely that any values of $\theta_1, \ldots, \theta_m$ will satisfy (1.27). A problem arises with (1.28) if the parameters are on very different scales, since if $m = 2$, and θ_1 is in the range (10, 100) and θ_2 in the range (0.001, 0.01), then (1.28) will virtually ignore the second parameter. This problem of scale will also affect those optimization methods which are not invariant with respect to scale changes. The obvious solution to this problem is to choose units for the parameter so that each has roughly the same magnitude. For more detailed comments about stopping criteria and scaling see Dennis and Schnabel (1983, Ch. 7).

1.5 CONSTRAINED MINIMIZATION

In the discussion in previous sections it has been implicitly assumed that the elements of the parameter vector, $\boldsymbol{\theta}$, are not subject to any constraints. This is not always the case, however, and problems do arise where we wish to minimize some objective function, $f(\boldsymbol{\theta})$, subject to various constraints on the parameters. Such constraints may be equalities, for example,

$$\theta_1^2 + \theta_2^2 + \theta_3^2 = 1 \tag{1.29}$$

or inequalities,

$$\theta_1 + \theta_2 + \theta_3 > 0 \tag{1.30}$$

Constraints on the parameters in statistical problems may arise for a number of reasons; the parameters may, for example, be variances which must be greater than zero, or proportions which must lie between zero and one.

The simplest method of dealing with constrained optimization problems is to reparametrize so that they become unconstrained. For example, if an original parameter is subject to constraints of the form

$$\theta_i > c_i \tag{1.31}$$

$$a_i < \theta_i < b_i \tag{1.32}$$

where a_i, b_i and c_i are constants, then defining a new parameter α_i as

$$\alpha_i^2 = \theta_i - c_i \tag{1.33}$$

$$\sin^2 \alpha_i = (\theta_i - a_i)/(b_i - a_i) \tag{1.34}$$

removes the constraints (1.31) and (1.32) and allows an unconstrained optimization for the parameter α_i. Particularly common in statistics is the situation where a_i and b_i in (1.32) are zero and unity and the parameter θ_i represents a proportion or probability. A commonly used transformation in this case is the *logistic*,

$$\alpha_i = \log \frac{\theta_i}{1-\theta_i} \tag{1.35}$$

More formal methods of dealing with constrained optimization problems such as *Lagrange multipliers* and *penalty functions* are described in Rao (1979, Ch. 7). It is important to emphasize, however, that many problems with simple constraints can be solved by unconstrained algorithms, because the constraints are satisfied by the unconstrained minimizer.

1.6 SUMMARY

Many problems in statistics may be formulated in terms of the minimization of some function with respect to a number of parameters. In most cases the equations for a minimum arising from (1.7) cannot be solved directly, and iterative procedures are needed. During the last two decades there have been major advances in such techniques and this has had considerable impact in many branches of statistics, as we shall attempt to describe in later chapters. The next two chapters concentrate on describing a number of commonly used minimization methods. It should be emphasized that they do not attempt to provide a comprehensive account of such techniques, only to provide a basis for a discussion of their use in a statistical context. Many excellent *detailed* accounts of optimization methods are available elsewhere, for example, Bunday (1984), Walsh (1975) and Rao (1979).

2

Direct search methods

2.1 INTRODUCTION

In this chapter we shall describe a number of *direct search methods* for minimization. Such methods do *not* require the explicit evaluation of any partial derivatives of the function being minimized, but instead rely solely on values of the function found during the iterative process. In some cases these function values are used to obtain numerical approximations to the derivatives of the objective function, in others they provide the basis for fitting low-order polynomials or surfaces to the function in the vicinity of the minimum. We first consider the minimization of a function of a single parameter, and then the multiparameter situation.

2.2 UNIVARIATE SEARCH METHODS

Search methods for minimizing a function of a single variable fall into two classes: those which specify an interval in which the minimum lies and those which specify the position of the minimum by a point approximating to it. In order to apply the former we shall assume that an initial interval known to contain the minimum is given and that the function is unimodal within this interval. With such methods we literally search for the minimum of the function in some interval $a < \theta < b$ by evaluating the function at chosen points in the interval. The alternative approach is to use a few function values evaluated at particular points to approximate the function by a simple polynomial, at least over a limited range of values. The position of the function minimum is then approximated by the position of the polynomial minimum, the latter being relatively simple to calculate. We begin with an example of the first approach followed by one of the second.

2.2.1 Fibonacci search

We suppose that the required minimum is known to be within the interval (θ_1, θ_2), and that two points, θ_3 and θ_4, are to be chosen within this interval so that

$$\theta_1 < \theta_3 < \theta_4 < \theta_2 \tag{2.1}$$

此为试读,需要完整PDF请访问: www.ertongbook.com