6 Constrained functions

6.1 Introduction

Up to now we have only asked smoothness of our functions, but in many situations the function that we estimate must also satisfy important side conditions, such as being strictly increasing. Unfortunately, our central idea of using a basis expansion can get us into trouble here. We saw in Chapters 4 and 5 that smoothing the height data often produced curves that did not increase everywhere and consequently had negative velocities. It is, in general, difficult to force functions defined by linear expansions to satisfy constraints such as being everywhere positive, monotone, and so forth.

In this chapter we consider four constrained estimation situations: functions which must be positive, those which must be strictly monotone, those whose values are probabilities, and probability density functions, which must be both positive and integrate to one. We will in each case redefine the original problem so that the function to be estimated is unconstrained. The idea of defining a constrained function by a differential equation will be introduced. In the density estimation case, it will be necessary to use a fitting criterion other than least squares.

6.2 Fitting positive functions

Data are often collected on functions that are strictly positive. The data themselves may be zero, but these zero values indicate only that the func-



Figure 6.1. The circles indicate squared residuals averaged across the 54 girls in the Berkeley growth study for the ages of observation. The solid line is a positive smooth of these data.

tion being estimated is small at that point. For example, within any year we are apt to observe many days without any rainfall, and there can be days where no rain is recorded in thirty years. But eventually it rains, at least in Canada. Counts of errors, as another example, can easily be zero, but eventually everybody makes a mistake.

We often need to use the data to estimate variances $\sigma^2(t)$ that vary over argument t, and a zero or negative variance estimate can cause all kinds of problems. We don't want a string of zero observed variances to translate into a zero estimate.

We can estimate a nonstationary variance function for the 54 females in the Berkeley growth data as follows. First, we smooth the height values assuming a constant variance. We opted for smoothing splines with a penalty on $(D^4x)^2$ and $\lambda = 1$. The squared residuals from these fits were averaged over cases, and Figure 6.1 displays the resulting averages. The measurement error is elevated in infancy and around the pubertal growth spurt, but the error variance estimates near the boundaries are certainly too small, probably due to the over-fitting that happens in these regions due to the lack of data.

The solid line in Figure 6.1 shows the fit using a version of a smoothing spline, to be described below, that is constrained to be positive, so as to avoid any estimated value $\hat{\sigma}^2(t)$ that is zero or negative. The fit indicates that the standard error of measurement is about seven millimeters in infancy, but is more like three millimeters in later years. We then re-smoothed

the data using a diagonal weight matrix \mathbf{W} containing its diagonal the reciprocals of the fitting variances $w_j = 1/\sigma^2(t_j)$. We could go on to iterate this process by re-estimating average squared residuals, and so on.

6.2.1 A positive smoothing spline

A positive smoothing function x can always be defined as the exponential of an unconstrained function W:

$$x(t) = e^{W(t)} , (6.1)$$

so that W is the logarithm of x. Other bases besides e for the logarithm are, of course, always appropriate; some of our clients may prefer the base 10 to aid interpretation.

Since W(t) can be positive or negative and is not in any other way constrained, it is reasonable to expand W in terms of a set of basis functions,

$$W(t) = \sum_{k} c_k \phi_k(t) , \qquad (6.2)$$

probably using a Fourier series for periodic data such as daily precipitation levels and B-spline expansions for non-periodic data such as the mean squared residuals for the growth data.

The roughness of a positive smoothing function x is defined as the roughness of its logarithm W, so that the roughness-penalized fitting criterion for positive smoothing, using the size of the second derivative, is

$$\text{PENSSE}_{\lambda}(W|\mathbf{y}) = [\mathbf{y} - e^{W(\mathbf{t})}]' \mathbf{W} [\mathbf{y} - e^{W(\mathbf{t})}]^2 + \lambda \int [D^2 W(t)]^2 dt . \quad (6.3)$$

A complication on the computational side is that we must now use numerical methods to minimize criterion (6.3) with respect to the coefficients c_k of the expansion. These methods iteratively decrease an initial estimate of W(t) until convergence is reached. However, because the exponential transform is only mildly nonlinear, these iterative methods usually converge rapidly, even from initial estimates far away from the final value. In fact, we find that starting with W = 0 works just fine in most circumstances. Keep in mind, however, that if the data are mostly zero in a region, and especially at the boundaries, the values of W(t) in that region will be poorly defined large negative numbers.

The positive smooth of the residuals in Figure 6.1 was obtained by using an order four B-spline expansion of W(t) with a knot located at each age of observation. The fit was made smooth by using as a roughness penalty the integrated squared derivative $D^2\sigma$ multiplied by $\lambda = 0.0001$.

114 6. Constrained functions

6.2.2 Representing a positive function by a differential equation

A differential equation expresses a relationship between a function and one or more of its derivatives, and is often an elegant way of describing functions with special structures.

For example, what does the notation e^{wt} , where w is here some fixed rate constant, really mean? We may say that the notation stands for a recipe for computing its value, namely the convergent infinite series

$$x(t) = \sum_{i=0}^{\infty} \frac{(wt)^i}{i} .$$
 (6.4)

But a recipe is not the same as a taste, and a computer program is not the same as the mathematical concept whose value it calculates. We might prefer a definition that tells us directly about an important property of e^{wt} . Here it is:

$$Dx(t) = w(t)x(t).$$
(6.5)

This is easier to remember, and a positive w evokes the image of a graph that increases more and more rapidly as the function gets larger and larger, that is, an image of explosive growth. Or, if w is negative, that the slope of x goes to zero as the function value x(t) goes to zero.

If w(t) is a function, the solution function x for the differential equation (6.5) is

$$x(t) = C \exp[\int_{t_0}^t w(u) \, du]$$
(6.6)

for some nonzero constant C and lower limit of integration t_0 . In the cleaner functional notation, $x = C \exp D^{-1} w$. If C > 0, then

$$x(t) = \exp[W(t)], \tag{6.7}$$

where

$$W(t) = \int_{t_0}^t w(u) \, du + \log C = D^{-1} w(t) + \log C.$$

In fact, our invocation of the infinite series (6.4) would not be correct in a wider functional sense; the recipe (6.4) only works when w is a constant and therefore $(D^{-1}w)(t) = wt$. Three lessons are therefore to be drawn:

- Going from scalar to functional notation can turn up some surprises.
- A differential equation can be a powerful and evocative way of defining a function.
- The solution to a differential equation is a *class of functions*, in this case corresponding to the arbtrary choice of constant *C*.

Perhaps the need for this little bit of mathematics will be less than apparent here. If so, ignore it, but do expect the differential equation theme to return again and again, and to become progressively more important.

6.3 Fitting strictly monotone functions

Now that we have the principle that smooth functions can result from transforming the constrained smoothing problem to one that is unconstrained, we are ready to look at the monotone smoothing of the growth data. A strictly monotone function has a strictly positive first derivative. The spline smoothing approach that has been used up to now has worked fine for ages up to about sixteen, but after that the estimated velocities have in many cases gone negative. We hope that preventing negative estimates of velocities can stabilize height, velocity and acceleration estimates at the adult end of the fitting interval.

6.3.1 Fitting the growth of a baby's tibia

Figure 6.2 displays a tough monotone smoothing problem. The data were collected by M. Hermanussen et al. (1998), who developed an instrument measuring bone lengths to within about 0.1 millimeters. They are the lengths of the tibia, the large bone in the lower leg, in a newborn baby measured daily for the first 40 days of its life. It is clear that growth at this age is not a smooth process; a few days of little growth are separated by the astonishing increase of two or more millimeters within twenty-four hours. The only way a conventional unconstrained smoother could avoid having negative slope would be to smooth so heavily that the data would be badly fit. We especially need to fit the data monotonically here in order to get a good estimate of the velocity of growth, displayed in Figure 6.3.

6.3.2 Expressing a strictly monotone function explicitly

The solution x to the strictly monotone smoothing problem is linked to positive function estimation in Section 6.2 since velocity Dx is now assumed to be positive. We can, therefore, use (6.1) by expressing Dx as the exponential of an unconstrained function W to obtain

$$Dx(t) = e^{W(t)}$$
 (6.8)

By integrating both sides of this equation, we obtain

$$x(t) = C + \int_{t_0}^t \exp[W(u)] \, du \,\,, \tag{6.9}$$

where C is a constant that will have to be estimated from the data.



Figure 6.2. The length of the tibia of a newborn infant over the first 40 days of life. The solid line is the fit to the data using a monotone smoothing spline.

6.3.3 Expressing a strictly monotone function as a differential equation

Again we can pass directly from differential equation (6.5) to the corresponding equation for monotone functions by substituting Dx for x:

$$D^2 x = w D x av{6.10}$$

Here function $w = D^2 x/Dx$, and is consequently the derivative of the logarithm of velocity, the log velocity always existing because velocity is positive.

This differential equation has the following general solution:

$$x(t) = C_0 + C_1 \int_{t_0}^t \exp[\int_{t_0}^v w(v) \, dv] \, du \;, \tag{6.11}$$

where C_1 is nonzero. This is the same equation as (6.9) if we substitute

$$W(u) = \int_{t_0}^{v} w(v) \, dv + \log C_1 = D^{-1} w(u) + \log C_1 \, .$$

Let us consider the role of function w. First, if w(t) = 0 for all t, we have the solution

$$x(t) = C_0 + C_1 t \; .$$



Figure 6.3. The estimated derivative or velocity of the data in Figure 6.2.

If w is a nonzero constant, then the solution becomes

$$x(t) = C_0 + C_1 e^{wt} .$$

Thus, linear functions are at the origin of a sort of one-dimensional functional coordinate system defined by varying function w, and exponential functions are contained within the same system. If w is a function, then the closer to zero it is at an argument value t, the more its local behavior around t will be linear. If C_1 is positive, then positive values of w(t)imply local exponential increase, and negative values imply an exponential approach locally to some asymptote.

We will be especially interested in the next chapter in strictly monotone functions, called *warping functions*, that monotonically transform a time interval [0, T] into itself. There the differential equation will reveal other neat properties.

6.4 The performance of spline smoothing revisited

In Chapter 5 we used direct spline smoothing to fit simulated growth data for girls. We now ask how monotone smoothing compares in performance with this direct smoothing. We simulated 1000 samples from the mean curve used in Section 5.5, but this time fit each curve with a monotone smooth, penalizing the variation in the third derivative of relative acceleration function W with a smoothing parameter of 0.1. This is roughly



Figure 6.4. The cross-hatched area shows point-wise 95% confidence limits obtained from monotone smoothing, and the other dashed lines are the limits obtained with direct smoothing that are shown in Figure 5.7.

equivalent to the the nature and amount of roughness penalization that we used for the direct spline smoothing.

The results are shown in Figure 6.4 as point-wise confidence limits around the true curve, along with the standard error estimates that we obtained previously. There is a great improvement in precision of estimation at later ages, where the monotonicity constraint acts as a powerful smoothing principle in its own right. There is also much improvement in precision in the childhood ages as well, where we very much need the extra fitting power in order to study the smaller "mid-spurts" that are often found there. We lose, though, in early childhood, where the steep slope on the acceleration function leaves a lot of room for variation, and where violating monotonicity is no problem for the direct smoothing estimate.

6.5 Fitting probability functions

It is often necessary to estimate the probability of an event happening as a function of time or some other continuum. Does, for example, the probability that a non-smoking worker will get lung cancer depend on the number of cigarettes smoked per hour in his workspace? This *probability function* with values p(n) is an example of a *dose response function* of the kind often estimated in pharmacokinetics and toxicology. A function p taking values on the interior of the unit interval (0, 1) can be neatly defined by a differential equation. The differential equation Dx = wxworked for nonzero functions because, by definition, x is never zero. In this case, what is never zero is p(t)[1 - p(t)]. Consequently, we can propose the nonlinear differential equation

$$Dp(t) = w(t)p(t)[1 - p(t)] . (6.12)$$

The equation implies that

$$w(t) = \frac{Dp(t)}{p(t)[1-p(t)]}$$

so that function w(t) is the slope of p at t relative to the variance of the binary variable with p(t) as its parameter.

The explicit solution to this equation is

$$p(t) = \frac{\exp[\int_{t_0}^t w(u) \, du]}{1 + \exp[\int_{t_0}^t w(u) \, du]} , \qquad (6.13)$$

and, defining

$$W(t) = \int_{t_0}^t w(u) \, du,$$

we have that

$$W(t) = \log\left[\frac{p(t)}{1 - p(t)}\right]$$

is the *log odds-ratio* function.

An example using of this formulation of the binomial smoothing problem can be found in Chapter 9 of Ramsay and Silverman (2002) and in Rossi, Wang and Ramsay (2002).

We would not normally choose to fit a set of frequency sample size pairs (f_j, N_j) by least squares estimation. Rather, we would use maximum likelihood estimation, or treat the model as a general linear model (GLM), which amounts to the same thing.

6.6 Estimating probability density functions

The estimation of a probability density function p describing the distribution of a set of sample values t_1, \ldots, t_N is perhaps one of the oldest functional estimation problems in statistics. A probability density function is positive, and therefore is a special case of (6.6), and thus of the form

$$p(t) = Ce^{W(t)}$$

120 6. Constrained functions

but with the additional restriction

$$\int p(t) \, dt = 1.$$

The constant C in (6.6) is therefore

$$C = 1/\big(\int e^{W(t)} \, dt\big)$$

Maximizing likelihood is the usual strategy for estimating a density function, Given N observed values t_k , we would in practice maximize the log likelihood

$$\ln L(W|\mathbf{c}) = \sum_{i}^{N} \log p(t_{i})$$

$$= \sum_{i}^{N} W(t_{i}) - N \ln \int e^{W(t)} dt$$

$$= \sum_{i}^{N} \mathbf{c}' \phi(t_{i}) - N \ln \int \exp[\mathbf{c}' \phi(t)] dt \qquad (6.14)$$

where $W(t) = \mathbf{c}' \boldsymbol{\phi}(t)$ for a vector $\boldsymbol{\phi}$ of K basis functions.

The roughness of the estimated density can always be controlled by the number K of basis functions, but the versatility of roughness penalties that we have already encountered suggests they might work better here, too. If we maximize the log likelihood, we will have to subtract the roughness penalty to control roughness.

Using the penalty

$$\operatorname{PEN}_3(W) = \int [D^3 W(t)]^2 dt$$

implies that the heavier the penalty, the more W will approximate a quadratic function and, consequently, the more density p will approach a normal or Gaussian density function (Silverman, 1982, 1986). Later in Chapter 18 it will be shown that there is a linear differential operator L corresponding to most of the textbook density functions such that penalizing the size of LW can permit us to smooth toward one of a wide range of default densities.

Figure 6.5 shows the probability density function for the log of daily rainfall at Prince Rupert, British Columbia, one of the rainiest places in North America, over the years 1960 through 1994. Even there, however, about a third of the days have a precipitation of 0.1 mm or less, and we used only the 7697 days having precipitation in excess of 0.1 mm. Sixteen B-spline basis functions of order five and equally spaced knots were used to expand W(t), and the size of the third derivative was penalized with $\lambda = 10^{-6}$. The distribution is rather negatively skewed, even after taking



Figure 6.5. The estimated probability density function for the log (base 10) of daily precipitation at Prince Rupert, British Columbia. Only the 7697 precipitations in excess of 0.1 mm were used.

a log transformation. The sharp peak in the density suggests that rainfall comes in two forms: a steady drizzle that leaves up to a centimeter of rain, and large violent storms that can dump more than 10 centimeters of rain. It is possible to show point-wise confidence regions for density estimates such as these, but with this many observations the limits are too close to the estimated curve to be worth plotting.

6.7 Functional data analysis of point processes

A point process is a sample of N times of events, t_1, \ldots, t_N . These times are usually taken relative to some time t_0 at which recording begins, and this can be taken without losing any generality as $t_0 = 0$.

There are two questions that are central to point processes:

- Given that an event has already taken place at time t_i , how probable is it that the next event will take place at a time at or near $t \ge t_i$?
- What is the relation of this probability to the events already observed? For example, how does the probability that the next event will be at t depend on the time $t_i - t_{i-1}$ between the last two events?

The simplest of point processes, the *homogeneous Poisson process*, answers these two questions in this way. First, let there be no relationship whatever between the time of the next event and the times of previous events. Second, the distribution of the time to the next event, that is $t - t_i$, is *exponential*. The distribution function and density function for an exponential distribution, respectively, are

$$F(t-t_i|\mu) = 1 - \exp[-\mu(t-t_i)]$$
 and $p(t-t_i|\mu) = \mu \exp[-\mu(t-t_i)]$, (6.15)

where the parameter μ is average number of events per unit time, and is often called the *intensity parameter* of the process. The exponential distribution is a model of perfect chaos in the sense that if you have waited already to time t for an event to occur, the distribution of how much longer you have to wait remains exponential. That is, you gain nothing by waiting. The larger μ , the shorter your average waiting time, which is $1/\mu$.

The likelihood $L(t_1, \ldots, t_N)$ of the sample of event times is, using $t_0 = 0$,

$$L(\mathbf{t}|\mu) = \prod_{i}^{N} p(t_i - t_{i-1}|\mu)$$

and the log likelihood is

$$\ln L(\mathbf{t}|\mu) = \sum_{i}^{N} [\ln \mu - \mu(t_{i} - t_{i-1})] = N \ln \mu - \mu t_{N}.$$
(6.16)

Consequently, the maximum likelihood estimate of μ is

$$\hat{\mu} = \frac{N}{t_N} \; ,$$

and it is interesting to note that the estimate depends only on the last observed time, and thus ignores previous event times.

Poisson processes, although well understood by statisticians, are usually much too simple to model real-life event times. For example, if you have waited twenty minutes for a bus, it is reasonable to assume that you won't have to wait much longer. Also, the probability of a particular waiting time is often not likely to remain constant, as (6.15) suggests; if you are waiting for a bus at 3 a.m., you can expect to wait longer than at 5 p.m.

We may decide to keep the assumption of independence of event times, but relax the assumption of a constant intensity parameter. Suppose, now, that intensity parameter μ is a function of time with values $\mu(t)$. The mathematically natural way to generalize (6.15) to this situation is

$$F(t - t_i|\mu) = 1 - \exp[-\int_{t_i}^t \mu(s) \, ds]$$

$$p(t - t_i|\mu) = \mu(t) \exp[-\int_{t_i}^t \mu(s) \, ds].$$
 (6.17)

6.8. Fitting a linear model with estimation of the density of residuals 123

This more general model reduces to the Poisson process when $\mu(s)$ is a constant. The log likelihood now becomes

$$\ln L(\mathbf{t}|\mu) = \sum_{i}^{N} \ln \mu(t_{i}) - \int_{0}^{t_{N}} \mu(s) \, ds.$$
(6.18)

The fact that $\mu(t)$ is nonnegative suggests that we should take the approach in Section 6.2 and use the exponentiated basis function expansion

$$\mu(t) = \exp[\mathbf{c}'\boldsymbol{\phi}(t)], \qquad (6.19)$$

where ϕ is a functional vector of K basis functions, and vector **c** contains the coefficients of the expansion. Substituting this into 6.18 gives us

$$\ln L(\mathbf{t}|\mu) = \sum_{i}^{N} \mathbf{c}' \boldsymbol{\phi}(t_i) - \int_{0}^{t_N} \exp[\mathbf{c}' \boldsymbol{\phi}(s)] \, ds.$$
 (6.20)

In order to compute the maximum likelihood estimate of the coefficients in \mathbf{c} , we need the derivative

$$D_{\mathbf{c}} \ln L(\mathbf{t}|\mu) = \sum_{i}^{N} \boldsymbol{\phi}(t_i) - \int_{0}^{t_N} \boldsymbol{\phi}(s) \exp[\mathbf{c}' \boldsymbol{\phi}(s)] \, ds.$$
(6.21)

Setting $D_{\mathbf{c}} \ln L$ to zero does not result in any simple solution for \mathbf{c} , and we must resort to numerical optimization methods to maximize (6.20).

If we compare the log likelihood in this situation with (6.14) for the log likelihood in the problem density estimation, the similarity is striking. The first term is the same, and the second term for densities involves multiplying by N, logging and then integrating to ∞ rather than just integrating to t_N . The problems are thus essentially the same except for relatively minor changes in the normalizing constraint.

Lupus is an autoimmune disease characterized by sudden flares in symptoms. Figure 6.6 shows the timings of 41 flares for a single patient over nearly 19 years, along with the estimated intensity function μ for these data. The intensity function reflects well the two periods when this patient was relatively free of flares, as well as the period of intense disease activity around year eight. The point-wise confidence limits, however, caution us that this amount of data does not pin down the intensity function especially well. These results were achieved using 13 order four B-splines with a roughness penalty on $D^2\mu$ multiplied by smoothing parameter 5.0.

6.8 Fitting a linear model with estimation of the density of residuals

Our default approach to fitting data has been to minimize the sum of squared residuals. This is tantamount to assuming that the residuals are



Figure 6.6. The times of flares in lupus symptoms for a single patient are indicated by vertical lines on the horizontal axis. The solid line is the intensity function μ for a nonhomogeneous Poisson process estimated from these data. The dashed lines indicate 95% point-wise confidence limits for the intensity function.



Figure 6.7. The estimated probability density functions of the residuals from fitting log (base 10) of daily precipitation at Churchill, Manitoba, and Vancouver, British Columbia.

normally distributed in the population of potential observations from which we have sampled.

Assuming normality can be risky, and especially if the true distribution has long tails. It would be safer if we could estimate the density of the residuals as well as the fit to the data by combining the methodology in Chapters 4 and 5 with the density estimation approach of Section 6.6. For example, if there are residuals many standard deviations from the curve, then allowing for this in estimating the fitted curve will give a measure of *robustness* to the fit. In other situations, the density of the residuals might be interesting on its own, as we saw in Figure 6.5.

Suppose that we have a vector \mathbf{y} of N dependent variable observations, and an N by p design matrix \mathbf{Z} available as a basis for a linear model for \mathbf{y} . In the curve-fitting situation, for example, \mathbf{Z} will contain the values $\phi_k(t_j)$ of the basis functions at the sampling points $t_i, i = 1..., N$. Let a vector \mathbf{e} of N residuals be of the form

$$\mathbf{e} = \mathbf{y} - \mathbf{Z}\mathbf{c},\tag{6.22}$$

where we expect to estimate the coefficients in p-vector \mathbf{c} .

In addition to estimating \mathbf{c} , however, we want an estimate of the density of the residuals \mathbf{e} , and we want to use that density in the estimation of \mathbf{c} . It will often be convenient to *standardize* the residuals prior to estimating the residuals by dividing them by a constant σ . If σ is a preliminary estimate of the standard deviation of the e_i 's, for example, this will normalize the interval over which the residuals are distributed to something not too far from [-3, 3]. Consequently, defining \mathbf{r} to be

$$\mathbf{r} = \mathbf{e}/\sigma,\tag{6.23}$$

we want an estimate p of the standardized residual density with values

$$p(r_i) = \frac{e^{W(r_i)}}{\int e^{W(u)} du}$$
$$= \frac{e^{W(\frac{y_i - \sum_j z_{ij} c_j}{\sigma})}}{\int e^{W(u)} du}, \qquad (6.24)$$

where, as with (6.2) but with a change of symbols,

$$W(r) = \sum_{k} b_k \psi_k(r).$$

The computational problem is now to maximize

$$\text{PENLIK}(W|\mathbf{b}, \mathbf{c}) = \sum_{i}^{N} \log p(t_i) = \sum_{i}^{N} W(t_i) - N \int e^{W(t)} dt - \int [LW(t)]^2 dt$$
(6.25)

with respect to both \mathbf{b} and \mathbf{c} . This criterion can, of course, be further augmented with a roughness penalty on the fit defined by \mathbf{Zc} .

Figure 6.7 shows the residual density functions estimated in fitting log precipitation for Churchill high up on Hudson's Bay and Vancouver on the lower Pacific coast. In both cases, we see some strong departures from normality. Both densities have much heavier negative tails than the normal



Figure 6.8. The estimated log (base 10) of daily precipitation at Churchill, Manitoba, and Vancouver, British Columbia. The heavy solid lines are the fits resulting from estimating the residual densities, and the lighter dashed lines were fits using least squares estimation.

distribution, and Vancouver in particular has a number of strong normalized residuals at about -4.5. Churchill has a shoulder on the negative side of its density, suggesting a rainy season and a dry season.

Figure 6.8 shows the two fitted precipitation curves, and in each case also the curve fit by least squares. Both curves fitted by the estimated density are smoother, and Vancouver's fit is also less perturbed by some large negative residuals in mid–summer, while the least squares fit was. Estimating the density in this case made the fit more robust.

6.9 Further notes and readings

For a general purpose introduction to nonparametric density estimation, see Silverman (1986). Scott (1992) also considers multivariate density estimation. Otherwise, there is a vast literature on this topic, and especially using kernel smoothing methods.

A thorough reference on point processes is Snyder and Miller (1991), but the older Cox and Lewis (1966) is rather more readable.