

however, observe the complete data. Rather, for the first n_1 individuals, we observe only whether or not outcome 1 occurs; for the next n_2 , whether or not outcome 2 occurs; and for the final $n_3 = n - n_1 - n_2$, whether or not outcome 3 occurs. Denote the incomplete data by $Y = (Y_1, Y_2, Y_3)'$, where Y_j has a binomial distribution with parameters (n_j, θ_j) , $j = 1, 2, 3$ independently. The complete data log likelihood is

$$L_X(\theta) = \sum_{j=1}^3 X_j \log \theta_j,$$

where $\theta_j \geq 0$ and $\theta_1 + \theta_2 + \theta_3 = 1$. It follows that

$$Q(\theta, \theta^{(0)}) = \sum_{j=1}^3 X_j^{(0)} \log \theta_j, \quad (\text{B.1})$$

where $X_j^{(0)} = E[X_j | Y; \theta^{(0)}]$ and

$$X_1^{(0)} = Y_1 + (n_2 - Y_2) \frac{\theta_1^{(0)}}{\theta_1^{(0)} + \theta_2^{(0)}} + (n_3 - Y_3) \frac{\theta_1^{(0)}}{\theta_1^{(0)} + \theta_3^{(0)}},$$

with similar expressions for $X_2^{(0)}$ and $X_3^{(0)}$. The M step involves maximization of the imputed complete data likelihood (B.1) to give $\theta_j^{(1)} = X_j^{(0)}/n$, $j = 1, 2, 3$. This leads to a very simple iteration that converges to the MLE based on the incomplete data. With data $Y = (15, 8, 12)$ and $(n_1, n_2, n_3) = (24, 15, 20)$ and an initial value of $\theta^{(0)} = (0.2, 0.5, 0.3)$, for example, the algorithm converges in eight iterations to the estimates $\hat{\theta} = (0.4011, 0.2522, 0.3467)$, which are accurate to four significant digits. The EM algorithm in this example can be shown to converge to the MLE from any starting value in the interior of the parameter space. Newton's method can also be applied directly to the incomplete log likelihood l_Y and is easily implemented. It converges more quickly and also gives, as a by-product, estimates of the covariance matrix of $\hat{\theta}$.

This example is for illustration only. The EM algorithm is most useful in much more complicated problems and often provides a simple and intuitively appealing algorithm. It can be very slow to converge, and as noted earlier, may not converge to the MLE. A more practical example of the usefulness of the algorithm is in the treatment of interval censored data in Section 3.9.1.

B.2 STEIJLES INTEGRATION

In combining discrete and continuous cases, or in writing score functions, estimating functions, test statistics, and estimators using counting process and martingale theory, we often encounter integrals of the form

$$\int_s^t f dG = \int_s^t f(u) dG(u).$$

This is called a *Stieltjes integral*, and in this section, we define the integral for the types of functions considered in this book and examine some of its properties. In this expression, the function f is referred to as the *integrand* and G as the *integrator*. For our purposes, the function G is a right-continuous function and differentiable at all but perhaps a finite or countable number of points, some of which may be jump discontinuities.

We take a rather informal approach to defining the Stieltjes integral. More complete definitions and derivations can be found, for example, in the books by Royden (1968) and Ash (1972). Consider first the case in which G is a nondecreasing function and suppose that $G: [0, \infty) \rightarrow [0, \infty)$ can be written as

$$G(t) = \int_0^t g(u) du + \sum_{0 < a_\ell \leq t} g_\ell, \quad 0 < t < \infty, \quad (\text{B.2})$$

where g is a nonnegative (Riemann) integrable function and $g(u) = G'(u)$ at all $u \in [0, \infty)$ except possibly at a finite or countable set of points, and $g_\ell > 0$ for all ℓ . Note that G has jump discontinuities at positive values a_1, a_2, \dots and $g_\ell = \Delta G(a_\ell) = G(a_\ell) - G(a_\ell^-)$, $\ell = 1, 2, \dots$. It is easy to see that G is a right-continuous function with left-hand limits. For a given function $f: [0, \infty) \rightarrow \mathfrak{R}$, the Stieltjes integral from s to t ($0 < s \leq t$) of f with respect to G is

$$\begin{aligned} \int_s^t f(u) dG(u) &= \int_{(s,t]} f(u) dG(u) \\ &= \int_s^t f(u) g(u) du + \sum_{a_\ell \in (s,t]} f(a_\ell) g_\ell, \end{aligned} \quad (\text{B.3})$$

provided that the integral and the sum in the final expression exist. Note that by convention, the upper point t , but not the lower point s , is included in the range of integration.

In taking (B.3) to define the Stieltjes integral, we are using results from Lebesgue integration. More specifically, (B.3) arises from interpreting (B.2) as a Lebesgue integral, where G defines the measure on the nonnegative real line. The general arguments leading to (B.3) involve successive approximations of f over the interval $(s, t]$ with step functions f_n which converge pointwise to f at all $u \in (s, t]$. The approximation f_n is a step function with n component steps and we define

$$\int_s^t f_n dG = \sum_{j=1}^{k_n} f_{nj} G(B_{nj}),$$

where $B_{nj} = \{u \in (s, t] : f_n(u) = f_{nj}\}$, f_{n1}, \dots, f_{nk_n} are the distinct values taken by f_n , and $G(B)$ is the measure assigned to the set B by G . It is possible to approximate any Borel measurable function f in this way. The Stieltjes (or Lebesgue-Stieltjes)

integral is then defined as

$$\int_s^t f dG = \lim_{n \rightarrow \infty} \int_s^t f_n dG,$$

which can be shown to be independent of the sequence of approximations f_n used and, in the case considered above, yields the result (B.3).

It is easily seen that $G(t) = \int_0^t dG(u)$, $t \geq 0$. Further, if $G(u)$ is continuous, then $\int_s^t f(u) dG(u) = \int_s^t f(u)g(u) du$. If G is a step function, so that $g'(u) = 0$, then $\int_s^t f(u) dG(u) = \sum_{a \in (s,t]} f(a)g(a)$. Integrals over open intervals can be obtained as limits. For example,

$$\int_{(s,t)} f(u) dG(u) = \lim_{v \rightarrow t^-} \int_{(s,v]} f(u) dG(u).$$

Similarly, we obtain the integral over $[s, t)$ by taking a limit as $v \rightarrow s^-$ of the integral over (v, t) .

In many instances, we consider integrals with respect to a function $G : [0, \infty) \rightarrow \mathfrak{R}$, where $G = G_1 - G_2$ and G_1 and G_2 are nondecreasing right-continuous functions of the type discussed above. If the integrals of f with respect to G_1 and G_2 exist, the Stieltjes integral of f with respect to G is defined to be

$$\begin{aligned} \int_s^t f dG &= \int_s^t f dG_1 - \int_s^t f dG_2 \\ &= \int_s^t f(u)[g_1(u) - g_2(u)] du + \sum_{a \in (s,t]} f(a) \Delta G(a). \end{aligned} \quad (\text{B.4})$$

Such integrals arise, in particular, as stochastic integrals with respect to martingales in Chapter 5 and elsewhere.

The usual formula for integration by parts can be extended to apply to Stieltjes integrals. Suppose that $F(t)$ and $G(t)$ are right-continuous functions of bounded variation on finite intervals and expressible as differences of nondecreasing right continuous functions as above. It can be shown that

$$\int_s^t F(u) dG(u) = F(u)G(u)|_s^t - \int_s^t G(u) dF(u) + \sum_{a \in (s,t]} \Delta F(a) \Delta G(a). \quad (\text{B.5})$$

Variations on this formula are also sometimes given. For example, it is sometimes useful to work with the left-continuous versions of F and G as the integrands,

$$\int_s^t F(u^-) dG(u) = F(u)G(u)|_s^t - \int_s^t G(u^-) dF(u) - \sum_{a \in (s,t]} \Delta F(a) \Delta G(a).$$

Since the sum on the right side of these expressions is zero when F and G have no jumps, these reduce to the usual formula for integration by parts in that special case. To establish these expressions, apply (B.4) to the left side of (B.5) to obtain

$$\begin{aligned} \int_s^t F(u)g(u) du + \sum_{a \in (s,t]} F(a) \Delta G(a) \\ &= \int_{(s,t]} [F(s) + \int_{(s,t]} dF(v)] g(u) du + \sum_{a \in (s,t]} F(a) \Delta G(a) \\ &= F(s)[G(t) - G(s)] + \int_{(s,t]} \int_{(v,t]} g(u) du dF(v) + \sum_{a \in (s,t]} F(a) \Delta G(a). \end{aligned}$$

Some inspection and calculation show that this reduces to

$$\begin{aligned} F(s)[G(t) - G(s)] + \int_{(s,t]} \int_{(v,t]} dG(u) dF(v) \\ &= F(s)[G(t) - G(s)] + \int_{(s,t]} [G(t) - G(v) + \Delta G(v)] dF(v) \end{aligned}$$

which reduces to the right side of (B.5).

B.3 SOFTWARE FOR FAILURE TIME ANALYSES

In the first edition of this book we included Fortran programs for applying the Cox regression model as an appendix, since there was a dearth of commercially available software for these analyses in 1980. In contrast, a review by Goldstein and Harrell (1998) lists 14 commercially available software packages that have substantial capabilities for failure time data analysis, including some form of Cox regression, and this list is not exhaustive. The 14 packages listed all included Kaplan-Meier estimators, and most allow the convenient fitting of parametric regression models. Several packages allow left truncation as well as right censoring and include some model building and model-checking procedures for Cox regression. Some provide various options for handling tied data, and some allow interval censoring. At present there does not appear to be commercially available software for regression parameter estimation in the semiparametric accelerated failure time model, presumably because numerical aspects have only recently been well addressed. Readers are probably best advised to contact the authors of recent papers (e.g., Jin et al., 2001) for information on software availability. On the other hand, several packages include censored data rank tests pertinent to the accelerated failure time class. Also, much of the material on cohort sampling and on missing or mis-measured covariate data (Chapter 11) has yet to be included in software packages, although some packages provide simple options for accommodating missing data.