9

Survival Analysis and the EM Algorithm

Survival analysis had its roots in governmental and actuarial statistics, spanning centuries of use in assessing life expectancies, insurance rates, and annuities. In the 20 years between 1955 and 1975, survival analysis was adapted by statisticians for application to biomedical studies. Three of the most popular post-war statistical methodologies emerged during this period: the Kaplan–Meier estimate, the log-rank test,[1] and Cox's proportional hazards model, the succession showing increased computational demands along with increasingly sophisticated inferential justification. A connection with one of Fisher's ideas on maximum likelihood estimation leads in the last section of this chapter to another statistical method that has "gone platinum," the EM algorithm.

9.1 Life Tables and Hazard Rates

An insurance company's *life table* appears in Table 9.1, showing its number of clients (that is, life insurance policy holders) by age, and the number of deaths during the past year in each age group,[2] for example five deaths among the 312 clients aged 59. The column labeled $\hat{S}$ is of great interest to the company's actuaries, who have to set rates for new policy holders. It is an estimate of survival probability: probability 0.893 of a person aged 30 (the beginning of the table) surviving past age 59, etc. $\hat{S}$ is calculated according to an ancient but ingenious algorithm.

Let X represent a typical lifetime, so

$$f_i = \Pr\{X = i\} \tag{9.1}$$

[1] Also known as the Mantel–Haenszel or Cochran–Mantel–Haenszel test.

[2] The insurance company is fictitious but the deaths y are based on the true 2010 rates for US men, per Social Security Administration data.

Table 9.1 *Insurance company life table; at each age, n = number of policy holders, y = number of deaths, $\hat{h}$ = hazard rate y/n, $\hat{S}$ = survival probability estimate* (9.6).

Age	n	y	$\hat{h}$	$\hat{S}$	Age	n	y	$\hat{h}$	$\hat{S}$
30	116	0	.000	1.000	60	231	1	.004	.889
31	44	0	.000	1.000	61	245	5	.020	.871
32	95	0	.000	1.000	62	196	5	.026	.849
33	97	0	.000	1.000	63	180	4	.022	.830
34	120	0	.000	1.000	64	170	2	.012	.820
35	71	1	.014	.986	65	114	0	.000	.820
36	125	0	.000	.986	66	185	5	.027	.798
37	122	0	.000	.986	67	127	2	.016	.785
38	82	0	.000	.986	68	127	5	.039	.755
39	113	0	.000	.986	69	158	2	.013	.745
40	79	0	.000	.986	70	100	3	.030	.723
41	90	0	.000	.986	71	155	4	.026	.704
42	154	0	.000	.986	72	92	1	.011	.696
43	103	0	.000	.986	73	90	1	.011	.689
44	144	0	.000	.986	74	110	2	.018	.676
45	192	2	.010	.976	75	122	5	.041	.648
46	153	1	.007	.969	76	138	8	.058	.611
47	179	1	.006	.964	77	46	0	.000	.611
48	210	0	.000	.964	78	75	4	.053	.578
49	259	2	.008	.956	79	69	6	.087	.528
50	225	2	.009	.948	80	95	4	.042	.506
51	346	1	.003	.945	81	124	6	.048	.481
52	370	2	.005	.940	82	67	7	.104	.431
53	568	4	.007	.933	83	112	12	.107	.385
54	1081	8	.007	.927	84	113	8	.071	.358
55	1042	2	.002	.925	85	116	12	.103	.321
56	1094	10	.009	.916	86	124	17	.137	.277
57	597	4	.007	.910	87	110	21	.191	.224
58	359	1	.003	.908	88	63	9	.143	.192
59	312	5	.016	.893	89	79	10	.127	.168

is the probability of dying at age i, and

$$S_i = \sum_{j \geq i} f_j = \Pr\{X \geq i\} \tag{9.2}$$

is the probability of surviving past age $i - 1$. The *hazard rate* at age i is by

definition

$$h_i = f_i/S_i = \Pr\{X = i \,|\, X \geq i\}, \tag{9.3}$$

the probability of dying at age i given survival past age $i-1$.

A crucial observation is that the probability S_{ij} of surviving past age j given survival past age $i-1$ is the product of surviving each intermediate year,

$$S_{ij} = \prod_{k=i}^{j}(1-h_k) = \Pr\{X > j \,|\, X \geq i\}; \tag{9.4}$$

first you have to survive year i, probability $1-h_i$; then year $i+1$, probability $1-h_{i+1}$, etc., up to year j, probability $1-h_j$. Notice that S_i (9.2) equals $S_{1,i-1}$.

$\hat{S}$ in Table 9.1 is an estimate of S_{ij} for $i = 30$. First, each h_i was estimated as the binomial proportion of the number of deaths y_i among the n_i clients,

$$\hat{h}_i = y_i/n_i, \tag{9.5}$$

and then we set

$$\hat{S}_{30,j} = \prod_{k=30}^{j}\left(1-\hat{h}_k\right). \tag{9.6}$$

The insurance company doesn't have to wait 50 years to learn the probability of a 30-year-old living past 80 (estimated to be 0.506 in the table). One year's data suffices.[3]

Hazard rates are more often described in terms of a *continuous* positive random variable T (often called "time"), having density function $f(t)$ and "reverse cdf," or survival function,

$$S(t) = \int_t^\infty f(x)\,dx = \Pr\{T \geq t\}. \tag{9.7}$$

The hazard rate

$$h(t) = f(t)/S(t) \tag{9.8}$$

satisfies

$$h(t)dt \doteq \Pr\{T \in (t, t+dt) | T \geq t\} \tag{9.9}$$

for $dt \to 0$, in analogy with (9.3). The analog of (9.4) is[†] †1

[3] Of course the estimates can go badly wrong if the hazard rates change over time.

$$\Pr\{T \geq t_1 | T \geq t_0\} = \exp\left\{-\int_{t_0}^{t_1} h(x)\, dx\right\} \tag{9.10}$$

so in particular the reverse cdf (9.7) is given by

$$S(t) = \exp\left\{-\int_0^t h(x)\, dx\right\}. \tag{9.11}$$

A one-sided exponential density

$$f(t) = (1/c)e^{-t/c} \qquad \text{for } t \geq 0 \tag{9.12}$$

has $S(t) = \exp\{-t/c\}$ and constant hazard rate

$$h(t) = 1/c. \tag{9.13}$$

The name "memoryless" is quite appropriate for density (9.12): having survived to any time t, the probability of surviving dt units more is always the same, about $1 - dt/c$, no matter what t is. If human lifetimes were exponential there wouldn't be old or young people, only lucky or unlucky ones.

9.2 Censored Data and the Kaplan–Meier Estimate

Table 9.2 reports the survival data from a randomized clinical trial run by **NCOG** (the Northern California Oncology Group) comparing two treatments for head and neck cancer: **Arm_A**, chemotherapy, versus **Arm_B**, chemotherapy plus radiation. The response for each patient is survival time in months. The + sign following some entries indicates *censored data*, that is, survival times known only to exceed the reported value. These are patients "lost to followup," mostly because the **NCOG** experiment ended with some of the patients still alive.

This is what the experimenters hoped to see of course, but it complicates the comparison. Notice that there is more censoring in **Arm_B**. In the absence of censoring we could run a simple two-sample test, maybe Wilcoxon's test, to see whether the more aggressive treatment of **Arm_B** was increasing the survival times. *Kaplan–Meier* curves provide a graphical comparison that takes proper account of censoring. (The next section describes an appropriate censored data two-sample test.) Kaplan–Meier curves have become familiar friends to medical researchers, a *lingua franca* for reporting clinical trial results.

Life table methods are appropriate for censored data. Table 9.3 puts the **Arm_A** results into the same form as the insurance study of Table 9.1, now

Table 9.2 *Censored survival times in days, from two arms of the* NCOG *study of head/neck cancer.*

Arm_A: Chemotherapy

7	34	42	63	64	74+	83	84	91
108	112	129	133	133	139	140	140	146
149	154	157	160	160	165	173	176	185+
218	225	241	248	273	277	279+	297	319+
405	417	420	440	523	523+	583	594	1101
1116+	1146	1226+	1349+	1412+	1417			

Arm_B: Chemotherapy+Radiation

37	84	92	94	110	112	119	127	130
133	140	146	155	159	169+	173	179	194
195	209	249	281	319	339	432	469	519
528+	547+	613+	633	725	759+	817	1092+	1245+
1331+	1557	1642+	1771+	1776	1897+	2023+	2146+	2297+

with the time unit being months. Of the 51 patients enrolled[4] in Arm_A, $y_1 = 1$ was observed to die in the first month after treatment; this left 50 at risk, $y_2 = 2$ of whom died in the second month; $y_3 = 5$ of the remaining 48 died in their third month after treatment, and one was lost to followup, this being noted in the l column of the table, leaving $n_4 = 40$ patients "at risk" at the beginning of month 5, etc.

$\hat{S}$ here is calculated as in (9.6) except starting at time 1 instead of 30. There is nothing wrong with this estimate, but binning the NCOG survival data by months is arbitrary. Why not go down to days, as the data was originally presented in Table 9.2? A Kaplan–Meier survival curve is the limit of life table survival estimates as the time unit goes to zero.

Observations z_i for censored data problems are of the form

$$z_i = (t_i, d_i), \tag{9.14}$$

where t_i equals the observed survival time while d_i indicates whether or not there was censoring,

$$d_i = \begin{cases} 1 & \text{if death observed} \\ 0 & \text{if death not observed} \end{cases} \tag{9.15}$$

[4] The patients were enrolled at different calendar times, as they entered the study, but for each patient "time zero" in the table is set at the beginning of his or her treatment.

Table 9.3 `Arm_A` *of the* `NCOG` *head/neck cancer study, binned by month; n = number at risk, y = number of deaths, l = lost to followup, h = hazard rate y/n; $\hat{S}$ = life table survival estimate.*

Month	n	y	l	h	$\hat{S}$	Month	n	y	l	h	$\hat{S}$
1	51	1	0	.020	.980	25	7	0	0	.000	.184
2	50	2	0	.040	.941	26	7	0	0	.000	.184
3	48	5	1	.104	.843	27	7	0	0	.000	.184
4	42	2	0	.048	.803	28	7	0	0	.000	.184
5	40	8	0	.200	.642	29	7	0	0	.000	.184
6	32	7	0	.219	.502	30	7	0	0	.000	.184
7	25	0	1	.000	.502	31	7	0	0	.000	.184
8	24	3	0	.125	.439	32	7	0	0	.000	.184
9	21	2	0	.095	.397	33	7	0	0	.000	.184
10	19	2	1	.105	.355	34	7	0	0	.000	.184
11	16	0	1	.000	.355	35	7	0	0	.000	.184
12	15	0	0	.000	.355	36	7	0	0	.000	.184
13	15	0	0	.000	.355	37	7	1	1	.143	.158
14	15	3	0	.200	.284	38	5	1	0	.200	.126
15	12	1	0	.083	.261	39	4	0	0	.000	.126
16	11	0	0	.000	.261	40	4	0	0	.000	.126
17	11	0	0	.000	.261	41	4	0	1	.000	.126
18	11	1	1	.091	.237	42	3	0	0	.000	.126
19	9	0	0	.000	.237	43	3	0	0	.000	.126
20	9	2	0	.222	.184	44	3	0	0	.000	.126
21	7	0	0	.000	.184	45	3	0	1	.000	.126
22	7	0	0	.000	.184	46	2	0	0	.000	.126
23	7	0	0	.000	.184	47	2	1	1	.500	.063
24	7	0	0	.000	.184						

(so $d_i = 0$ corresponds to a + in Table 9.2). Let

$$t_{(1)} < t_{(2)} < t_{(3)} < \ldots < t_{(n)} \tag{9.16}$$

denote the *ordered* survival times,[5] censored or not, with corresponding indicator $d_{(k)}$ for $t_{(k)}$. The *Kaplan–Meier estimate* for survival probability †2 $S_{(j)} = \Pr\{X > t_{(j)}\}$ is then† the life table estimate

$$\hat{S}_{(j)} = \prod_{k \le j} \left(\frac{n-k}{n-k+1} \right)^{d_{(k)}}. \tag{9.17}$$

[5] Assuming no ties among the survival times, which is convenient but not crucial for what follows.

$\hat{S}$ jumps downward at death times t_j, and is constant between observed deaths.

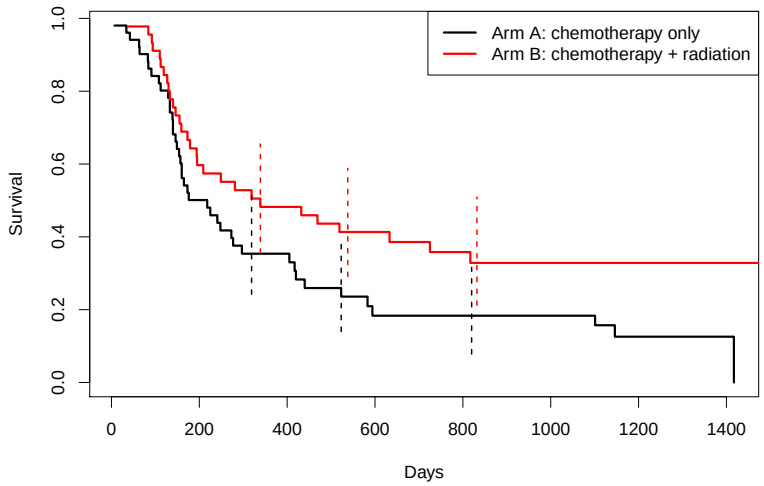

Figure 9.1 `NCOG` Kaplan–Meier survival curves; lower `Arm_A` (chemotherapy only); upper `Arm_B` (chemotherapy+radiation). Vertical lines indicate approximate 95% confidence intervals.

The Kaplan–Meier curves for both arms of the `NCOG` study are shown in Figure 9.1. `Arm_B`, the more aggressive treatment, looks better: its 50% survival estimate occurs at 324 days, compared with 182 days for `Arm_A`. The answer to the inferential question—is `B` really better than `A` or is this just random variability?—is less clear-cut.

The accuracy of $\hat{S}_{(j)}$ can be estimated from Greenwood's formula† for its standard deviation (now back in life table notation), †3

$$\text{sd}\left(\hat{S}_{(j)}\right) = \hat{S}_{(j)}\left[\sum_{k\le j}\frac{y_k}{n_k(n_k-y_k)}\right]^{1/2}. \tag{9.18}$$

The vertical bars in Figure 9.1 are approximate 95% confidence limits for the two curves based on Greenwood's formula. They overlap enough to cast doubt on the superiority of `Arm_B` at any one choice of "days," but the two-sample test of the next section, which compares survival at all timepoints, will provide more definitive evidence.

Life tables and the Kaplan–Meier estimate seem like a textbook example of frequentist inference as described in Chapter 2: a useful probabilistic

result is derived (9.4), and then implemented by the plug-in principle (9.6). There is more to the story though, as discussed below.

Life table curves are nonparametric, in the sense that no particular relationship is assumed between the hazard rates h_i. A parametric approach can greatly improve the curves' accuracy.† Reverting to the life table form of Table 9.3, we assume that the death counts y_k are independent binomials,

†4

$$y_k \overset{\text{ind}}{\sim} \text{Bi}(n_k, h_k), \tag{9.19}$$

and that the logits $\lambda_k = \log\{h_k/(1-h_k)\}$ satisfy some sort of regression equation

$$\boldsymbol{\lambda} = \boldsymbol{X}\alpha, \tag{9.20}$$

as in (8.22). A cubic regression for instance would set $x_k = (1, k, k^2, k^3)'$ for the kth row of $\boldsymbol{X}$, with $\boldsymbol{X}$ 47×4 for Table 9.3.

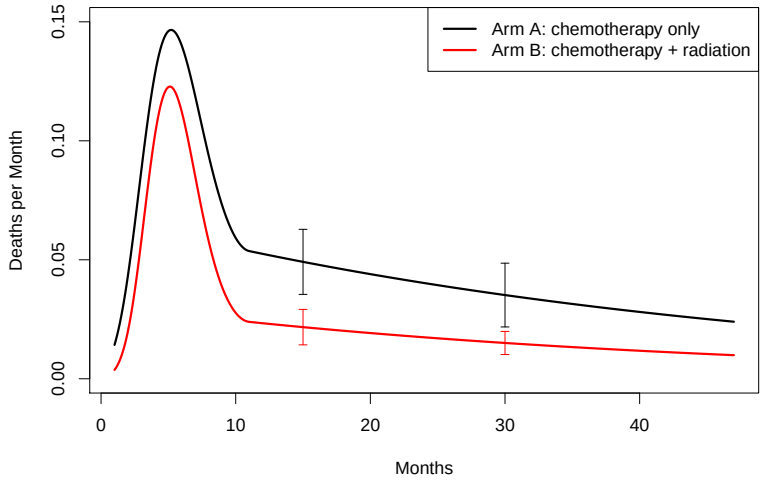

Figure 9.2 Parametric hazard rate estimates for the `NCOG` study. `Arm A`, black curve, has about 2.5 times higher hazard than `Arm B` for all times more than a year after treatment. Standard errors shown at 15 and 30 months.

The parametric hazard-rate estimates in Figure 9.2 were instead based on a "cubic-linear spline,"

$$x_k = \left(1, k, (k-11)_+^2, (k-11)_+^3\right)', \tag{9.21}$$

where $(k-11)_+$ equals $k-11$ for $k \le 11$, and 0 for $k \ge 11$. The vector

$\boldsymbol{\lambda} = \boldsymbol{X}\alpha$ describes a curve that is cubic for $k \leq 11$, linear for $k \geq 11$, and joined smoothly at 11. The logistic regression maximum likelihood estimate $\hat{\alpha}$ produced hazard rate curves

$$\hat{h}_k = 1 \Big/ \left(1 + e^{-x_k'\hat{\alpha}}\right) \tag{9.22}$$

as in (8.8). The black curve in Figure 9.2 traces $\hat{h}_k$ for **Arm_A**, while the red curve is that for **Arm_B**, fit separately.

Comparison in terms of hazard rates is more informative than the survival curves of Figure 9.1. Both arms show high initial hazards, peaking at five months, and then a long slow decline.[6] **Arm_B** hazard is always below **Arm_A**, in a ratio of about 2.5 to 1 after the first year. Approximate 95% confidence limits, obtained as in (8.30), don't overlap, indicating superiority of **Arm_B** at 15 and 30 months after treatment.

In addition to its frequentist justification, survival analysis takes us into the Fisherian realm of conditional inference, Section 4.3. The y_k's in model (9.19) are considered *conditionally* on the n_k's, effectively treating the n_k values in Table 9.3 as *ancillaries*, that is as fixed constants, by themselves containing no statistical information about the unknown hazard rates. We will examine this tactic more carefully in the next two sections.

9.3 The Log-Rank Test

A randomized clinical trial, interpreted by a two-sample test, remains the gold standard of medical experimentation. Interpretation usually involves Student's two-sample t-test or its nonparametric cousin Wilcoxon's test, but neither of these is suitable for censored data. The *log-rank test* † †5 employs an ingenious extension of life tables for the nonparametric two-sample comparison of censored survival data.

Table 9.4 compares the results of the **NCOG** study for the first six months[7] after treatment. At the beginning[8] of month 1 there were 45 patients "at risk" in **Arm_B**, none of whom died, compared with 51 at risk and 1 death in **Arm_A**. This left 45 at risk in **Arm_B** at the beginning of month 2, and 50 in **Arm_A**, with 1 and 2 deaths during the month respectively. (Losses

[6] The cubic–linear spline (9.21) is designed to show more detail in the early months, where there is more available patient data and where hazard rates usually change more quickly.

[7] A month is defined here as 365/12=30.4 days.

[8] The "beginning of month 1" is each patient's initial treatment time, at which all 45 patients ever enrolled in **Arm_B** were at risk, that is, available for observation.

Table 9.4 *Life table comparison for the first six months of the* NCOG *study. For example, at the beginning of the sixth month after treatment, there were 33 remaining* Arm_B *patients, of whom 4 died during the month, compared with 32 at risk and 7 dying in* Arm_A*. The conditional expected number of deaths in* Arm_A*, assuming the null hypothesis of equal hazard rates in both arms, was 5.42, using expression* (9.24).

Month	Arm_B		Arm_A		Expected number Arm_A deaths
	At risk	Died	At risk	Died	
1	45	0	51	1	.53
2	45	1	50	2	1.56
3	44	1	48	5	3.13
4	43	5	42	2	3.46
5	38	5	40	8	6.67
6	33	4	32	7	5.42

to followup were assumed to occur at the *end* of each month; there was 1 such at the end of month 3, reducing the number at risk in Arm_A to 42 for month 4.)

The month 6 data is displayed in two-by-two tabular form in Table 9.5, showing the notation used in what follows: n_A for the number at risk in Arm_A, n_d for the number of deaths, etc.; y indicates the number of Arm_A deaths. If the marginal totals n_A, n_B, n_d, and n_s are given, then y determines the other three table entries by subtraction, so we are not losing any information by focusing on y.

Table 9.5 *Two-by-two display of month-6 data for the* NCOG *study. E is the expected number of* Arm_A *deaths assuming the null hypothesis of equal hazard rates (last column of Table 9.4).*

	Died	Survived	
Arm_A	$y = 7$ $E = 5.42$	25	$n_A = 32$
Arm_B	4	29	$n_B = 33$
	$n_d = 11$	$n_s = 54$	$n = 65$

Consider the null hypothesis that the hazard rates (9.3) for month 6 are

the same in **Arm_A** and **Arm_B**,

$$H_0(6) : h_{A6} = h_{B6}. \tag{9.23}$$

Under $H_0(6)$, y has mean E and variance V,

$$\begin{aligned} E &= n_A n_d / n \\ V &= n_A n_B n_d n_s \big/ \left[n^2(n-1)\right], \end{aligned} \tag{9.24}$$

as calculated according to the *hypergeometric distribution*.† $E = 5.42$ and †6
$V = 2.28$ in Table 9.5.

We can form a two-by-two table for each of the $N = 47$ months of the **NCOG** study, calculating y_i, E_i, and V_i for month i. The log-rank statistic Z is then defined to be† †7

$$Z = \sum_{i=1}^{N} (y_i - E_i) \Bigg/ \left(\sum_{i=1}^{N} V_i\right)^{1/2}. \tag{9.25}$$

The idea here is simple but clever. Each month we test the null hypothesis of equal hazard rates

$$H_0(i) : h_{Ai} = h_{Bi}. \tag{9.26}$$

The numerator $y_i - E_i$ has expectation 0 under $H_0(i)$, but, if h_{Ai} is greater than h_{Bi}, that is, if treatment B is superior, then the numerator has a positive expectation. Adding up the numerators gives us power to detect a general superiority of treatment B over A, against the null hypothesis of equal hazard rates, $h_{Ai} = h_{Bi}$ for all i.

For the **NCOG** study, binned by months,

$$\sum_{i=1}^{N} y_i = 42, \quad \sum_{i=1}^{N} E_i = 32.9, \quad \sum_{i=1}^{N} V_i = 16.0, \tag{9.27}$$

giving log-rank test statistic

$$Z = 2.27. \tag{9.28}$$

Asymptotic calculations based on the central limit theorem suggest

$$Z \doteq \mathcal{N}(0, 1) \tag{9.29}$$

under the null hypothesis that the two treatments are equally effective, i.e., that $h_{Ai} = h_{Bi}$ for $i = 1, 2, \ldots, N$. In the usual interpretation, $Z = 2.27$ is significant at the one-sided 0.012 level, providing moderately strong evidence in favor of treatment B.

An impressive amount of inferential guile goes into the log-rank test.

1 Working with hazard rates instead of densities or cdfs is essential for survival data.
2 Conditioning at each period on the numbers at risk, n_A and n_B in Table 9.5, finesses the difficulties of censored data; censoring only changes the at-risk numbers in future periods.
3 Also conditioning on the number of deaths and survivals, n_d and n_s in Table 9.5, leaves only the *univariate* statistic y to interpret at each period, which is easily done through the null hypothesis of equal hazard rates (9.26).
4 Adding the discrepancies $y_i - E_i$ in the numerator of (9.25) (rather than say, adding the individual Z values $Z_i = (y_i - E_i)/V_i^{1/2}$, or adding the Z_i^2 values) accrues power for the natural alternative hypothesis "$h_{Ai} > h_{Bi}$ for all i," while avoiding destabilization from small values of V_i.

Each of the four tactics had been used separately in classical applications. Putting them together into the log-rank test was a major inferential accomplishment, foreshadowing a still bigger step forward, the *proportional hazards model*, our subject in the next section.

Conditional inference takes on an aggressive form in the log-rank test. Let $\boldsymbol{D}_i$ indicate all the data except y_i available at the end of the ith period. For month 6 in the NCOG study, $\boldsymbol{D}_6$ includes all data for months 1–5 in Table 9.4, and the marginals n_A, n_B, n_d, and n_s in Table 9.5, but not the y value for month 6. The key assumption is that, under the null hypothesis of equal hazard rates (9.26),

$$y_i | \boldsymbol{D}_i \overset{\text{ind}}{\sim} (E_i, V_i), \tag{9.30}$$

"ind" here meaning that the y_i's can be treated as independent quantities with means and variances (9.24). In particular, we can add the variances V_i to get the denominator of (9.25). (A "partial likelihood" argument, described in the endnotes, justifies adding the variances.)

The purpose of all this Fisherian conditioning is to simplify the inference: the conditional distribution $y_i | \boldsymbol{D}_i$ depends only on the hazard rates h_{Ai} and h_{Bi}; "nuisance parameters," relating to the survival times and censoring mechanism of the data in Table 9.2, are hidden away. There is a price to pay in testing power, though usually a small one. The lost-to-followup values l in Table 9.3 have been ignored, even though they might contain useful information, say if all the early losses occurred in one arm.

9.4 The Proportional Hazards Model

The Kaplan–Meier estimator is a one-sample device, dealing with data coming from a single distribution. The log-rank test makes two-sample comparisons. *Proportional hazards* ups the ante to allow for a full regression analysis of censored data. Now the individual data points z_i are of the form

$$z_i = (c_i, t_i, d_i), \tag{9.31}$$

where t_i and d_i are observed survival time and censoring indicator, as in (9.14)–(9.15), and c_i is a known $1 \times p$ vector of covariates whose effect on survival we wish to assess. Both of the previous methods are included here: for the log-rank test, c_i indicates treatment, say c_i equals 0 or 1 for `Arm_A` or `Arm_B`, while c_i is absent for Kaplan–Meier.

Table 9.6 *Pediatric cancer data, first 20 of 1620 children.* `Sex` *1 = male, 2 = female;* `race` *1 = white, 2 = nonwhite;* `age` *in years;* `entry` *= calendar date of entry in days since July 1, 2001;* `far` *= home distance from treatment center in miles;* `t` *= survival time in days;* `d` *= 1 if death observed, 0 if not.*

sex	race	age	entry	far	t	d
1	1	2.50	710	108	325	0
2	1	10.00	1866	38	1451	0
2	2	18.17	2531	100	221	0
2	1	3.92	2210	100	2158	0
1	1	11.83	875	78	760	0
2	1	11.17	1419	0	168	0
2	1	5.17	1264	28	2976	0
2	1	10.58	670	120	1833	0
1	1	1.17	1518	73	131	0
2	1	6.83	2101	104	2405	0
1	1	13.92	1239	0	969	0
1	1	5.17	518	117	1894	0
1	1	2.50	1849	99	193	1
1	1	.83	2758	38	1756	0
2	1	15.50	2004	12	682	0
1	1	17.83	986	65	1835	0
2	1	3.25	1443	58	2993	0
1	1	10.75	2807	42	1616	0
1	2	18.08	1229	23	1302	0
2	2	5.83	2727	23	174	1

Medical studies regularly produce data of form (9.31). An example, the *pediatric cancer* data, is partially listed in Table 9.6. The first 20 of $n = 1620$ cases are shown. There are five explanatory covariates (defined in the table's caption): **sex**, **race**, **age** at entry, calendar date of **entry** into the study, and **far**, the distance of the child's home from the treatment center. The response variable t is survival in days from time of treatment until death. Happily, only 160 of the children were observed to die ($d = 1$). Some left the study for various reasons, but most of the $d = 0$ cases were those children still alive at the end of the study period. Of particular interest was the effect of **far** on survival. We wish to carry out a regression analysis of this heavily censored data set.

The proportional hazards model assumes that the hazard rate $h_i(t)$ for the ith individual (9.8) is

$$h_i(t) = h_0(t)e^{c_i'\beta}. \tag{9.32}$$

Here $h_0(t)$ is a baseline hazard (which we need not specify) and β is an unknown p-parameter vector we want to estimate. For concise notation, let

$$\theta_i = e^{c_i'\beta}; \tag{9.33}$$

model (9.32) says that individual i's hazard is a constant nonnegative factor θ_i times the baseline hazard. Equivalently, from (9.11), the ith survival function $S_i(t)$ is a power of the baseline survival function $S_0(t)$,

$$S_i(t) = S_0(t)^{\theta_i}. \tag{9.34}$$

Larger values of θ_i lead to more quickly declining survival curves, i.e., to worse survival (as in (9.11)).

Let J be the number of observed deaths, $J = 160$ here, occurring at times

$$T_{(1)} < T_{(2)} < \ldots < T_{(J)}, \tag{9.35}$$

again for convenience assuming no ties.[9] Just before time $T_{(j)}$ there is a *risk set* of individuals still under observation, whose indices we denote by $\mathcal{R}_j$,

$$\mathcal{R}_j = \{i : t_i \geq T_{(j)}\}. \tag{9.36}$$

Let i_j be the index of the individual observed to die at time $T_{(j)}$. The key to proportional hazards regression is the following result.

[9] More precisely, assuming only one event, a death, occurred at $T_{(j)}$, with none of the other individuals being lost to followup at exact time $T_{(j)}$.

Lemma † *Under the proportional hazards model* (9.32)*, the conditional probability, given the risk set* $\mathcal{R}_j$*, that individual* i *in* $\mathcal{R}_j$ *is the one observed to die at time* $T_{(j)}$ *is* †8

$$\Pr\{i_j = i | \mathcal{R}_j\} = e^{c_i'\beta} \Big/ \sum_{k \in \mathcal{R}_j} e^{c_k'\beta}. \tag{9.37}$$

To put it in words, given that one person dies at time $T_{(j)}$*, the probability it is individual* i *is proportional to* $\exp(c_i'\beta)$*, among the set of individuals at risk.*

For the purpose of estimating the parameter vector β in model (9.32), we multiply factors (9.37) to form the *partial likelihood*

$$L(\beta) = \prod_{j=1}^{J} \left(e^{c_{i_j}'\beta} \Big/ \sum_{k \in \mathcal{R}_j} e^{c_k'\beta} \right). \tag{9.38}$$

$L(\beta)$ is then treated as an ordinary likelihood function, yielding an approximately unbiased MLE-like estimate

$$\hat{\beta} = \arg\max_{\beta} \{L(\beta)\}, \tag{9.39}$$

with an approximate covariance obtained from the second-derivative matrix of $l(\beta) = \log L(\beta)$,† as in Section 4.3, †9

$$\hat{\beta} \doteadot \left(\beta, \left[-\ddot{l}\left(\hat{\beta}\right)\right]^{-1}\right). \tag{9.40}$$

Table 9.7 shows the proportional hazards analysis of the pediatric cancer data, with the covariates **age**, **entry**, and **far** standardized to have mean 0 and standard deviation 1 for the 1620 cases.[10] Neither **sex** nor **race** seems to make much difference. We see that **age** is a mildly significant factor, with older children doing better (i.e., the estimated regression coefficient is negative). However, the dramatic effects are date of **entry** and **far**. Individuals who entered the study later survived longer—perhaps the treatment protocol was being improved—while children living farther away from the treatment center did worse.

Justification of the partial likelihood calculations is similar to that for the log-rank test, but there are some important differences, too: the proportional hazards model is semiparametric ("semi" because we don't have to specify $h_0(t)$ in (9.32)), rather than nonparametric as before; and the

[10] Table 9.7 was obtained using the **R** program **coxph**.

Table 9.7 *Proportional hazards analysis of pediatric cancer data (`age`, `entry` and `far` standardized). `Age` significantly negative, older children doing better; `entry` very significantly negative, showing hazard rate declining with calendar date of entry; `far` very significantly positive, indicating worse results for children living farther away from the treatment center. Last two columns show limits of approximate 95% confidence intervals for* $\exp(\beta)$.

	β	sd	z-value	p-value	$\exp(\beta)$	Lower	Upper
`sex`	−.023	.160	−.142	.887	.98	.71	1.34
`race`	.282	.169	1.669	.095	1.33	.95	1.85
`age`	−.235	.088	−2.664	.008	.79	.67	.94
`entry`	−.460	.079	−5.855	.000	.63	.54	.74
`far`	.296	.072	4.117	.000	1.34	1.17	1.55

emphasis on likelihood has increased the Fisherian nature of the inference, moving it further away from pure frequentism. Still more Fisherian is the emphasis on likelihood inference in (9.38)–(9.40), rather than the direct frequentist calculations of (9.24)–(9.25).

The conditioning argument here is less obvious than that for the Kaplan–Meier estimate or the log-rank test. Has its convenience possibly come at too high a price? In fact it can be shown that inference based on the partial likelihood is highly efficient, assuming of course the correctness of the proportional hazards model (9.32).

9.5 Missing Data and the EM Algorithm

Censored data, the motivating factor for survival analysis, can be thought of as a special case of a more general statistical topic, *missing data*. What's missing, in Table 9.2 for example, are the actual survival times for the + cases, which are known only to exceed the tabled values. If the data were *not* missing, we could use standard statistical methods, for instance Wilcoxon's test, to compare the two arms of the `NCOG` study. The EM algorithm is an iterative technique for solving missing-data inferential problems using only standard methods.

A missing-data situation is shown in Figure 9.3: $n = 40$ points have been independently sampled from a bivariate normal distribution (5.12),

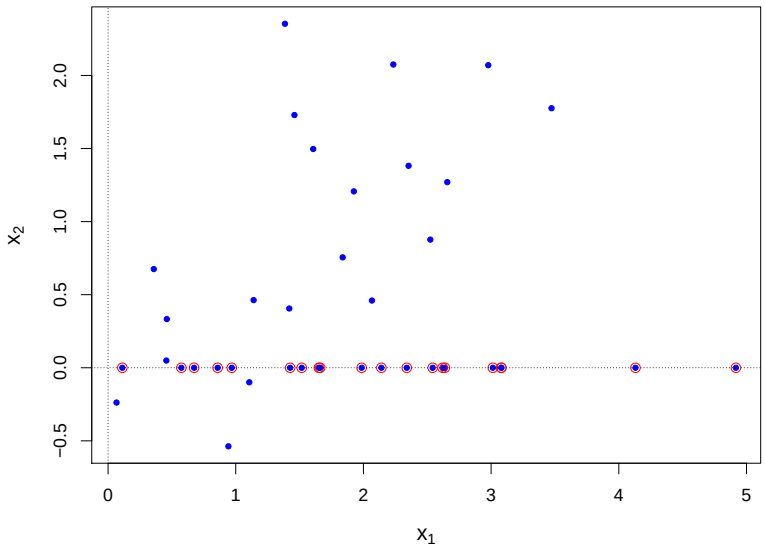

Figure 9.3 Forty points from a bivariate normal distribution, the last 20 with x_2 missing (circled).

means (μ_1, μ_2), variances (σ_1^2, σ_2^2), and correlation ρ,

$$\begin{pmatrix} x_{1i} \\ x_{2i} \end{pmatrix} \overset{\text{ind}}{\sim} \mathcal{N}_2 \left(\begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}, \begin{pmatrix} \sigma_1^2 & \sigma_1\sigma_2\rho \\ \sigma_1\sigma_2\rho & \sigma_2^2 \end{pmatrix} \right). \tag{9.41}$$

However, the second coordinates of the last 20 points have been lost. These are represented by the circled points in Figure 9.3, with their x_2 values arbitrarily set to 0.

We wish to find the maximum likelihood estimate of the parameter vector $\theta = (\mu_1, \mu_2, \sigma_1, \sigma_2, \rho)$. The standard maximum likelihood estimates

$$\begin{aligned} \hat{\mu}_1 &= \sum_{i=1}^{40} x_{1i}/40, \quad \hat{\mu}_2 = \sum_{i=1}^{40} x_{2i}/40, \\ \hat{\sigma}_1 &= \left[\sum_{i=1}^{40} (x_{1i} - \hat{\mu}_1)^2 /40 \right]^{1/2}, \quad \hat{\sigma}_2 = \left[\sum_{i=1}^{40} (x_{2i} - \hat{\mu}_2)^2 /40 \right]^{1/2}, \\ \hat{\rho} &= \left[\sum_{i=1}^{40} (x_{1i} - \hat{\mu}_1)(x_{2i} - \hat{\mu}_2) /40 \right] \Big/ (\hat{\sigma}_1 \hat{\sigma}_2), \end{aligned} \tag{9.42}$$

are unavailable for μ_2, σ_2, and ρ because of the missing data.

The EM algorithm begins by filling in the missing data in some way, say by setting $x_{2i} = 0$ for the 20 missing values, giving an artificially complete data set $data^{(0)}$. Then it proceeds as follows.

- The standard method (9.42) is applied to the filled-in $data^{(0)}$ to produce $\hat{\theta}^{(0)} = (\hat{\mu}_1^{(0)}, \hat{\mu}_2^{(0)}, \hat{\sigma}_1^{(0)}, \hat{\sigma}_2^{(0)}, \hat{\rho}^{(0)})$; this is the M ("maximizing") step.[11]
- Each of the missing values is replaced by its conditional expectation (assuming $\theta = \hat{\theta}^{(0)}$) given the nonmissing data; this is the E ("expectation") step. In our case the missing values x_{2i} are replaced by

$$\hat{\mu}_2^{(0)} + \hat{\rho}^{(0)} \frac{\hat{\sigma}_2^{(0)}}{\hat{\sigma}_1^{(0)}} \left(x_{1i} - \hat{\mu}_1^{(0)} \right). \tag{9.43}$$

- The E and M steps are repeated, at the jth stage giving a new artificially complete data set $data^{(j)}$ and an updated estimate $\hat{\theta}^{(j)}$. The iteration stops when $\|\hat{\theta}^{(j)+1} - \hat{\theta}^{(j)}\|$ is suitably small.

Table 9.8 shows the EM algorithm at work on the bivariate normal example of Figure 9.3. In exponential families the algorithm is guaranteed to converge to the MLE $\hat{\theta}$ based on just the observed data $\boldsymbol{o}$; moreover, the likelihood $f_{\hat{\theta}^{(j)}}(\boldsymbol{o})$ increases with every step j. (The convergence can be sluggish, as it is here for $\hat{\sigma}_2$ and $\hat{\rho}$.)

The EM algorithm ultimately derives from the *fake-data principle*, a property of maximum likelihood estimation going back to Fisher that can †10 only briefly be summarized here.[†] Let $\boldsymbol{x} = (\boldsymbol{o}, \boldsymbol{u})$ represent the "complete data," of which $\boldsymbol{o}$ is observed while $\boldsymbol{u}$ is unobserved or missing. Write the density for $\boldsymbol{x}$ as

$$f_\theta(\boldsymbol{x}) = f_\theta(\boldsymbol{o})\, f_\theta(\boldsymbol{u}|\boldsymbol{o}), \tag{9.44}$$

and let $\hat{\theta}(\boldsymbol{o})$ be the MLE of θ based just on $\boldsymbol{o}$.

Suppose we now generate simulations of $\boldsymbol{u}$ by sampling from the conditional distribution $f_{\hat{\theta}(\boldsymbol{o})}(\boldsymbol{u}|\boldsymbol{o})$,

$$\boldsymbol{u}^{*k} \sim f_{\hat{\theta}(\boldsymbol{o})}(\boldsymbol{u}|\boldsymbol{o}) \qquad \text{for } k = 1, 2, \ldots, K \tag{9.45}$$

(the stars indicating creation by the statistician and not by observation), giving fake complete-data values $\boldsymbol{x}^{*k} = (\boldsymbol{o}, \boldsymbol{u}^{*k})$. Let

$$data^* = \{\boldsymbol{x}^{*1}, \boldsymbol{x}^{*2}, \ldots, \boldsymbol{x}^{*K}\}, \tag{9.46}$$

[11] In this example, $\hat{\mu}_1^{(0)}$ and $\hat{\sigma}_1^{(0)}$ are available as the complete-data estimates in (9.42), and, as in Table 9.8, stay the same in subsequent steps of the algorithm.

Table 9.8 *EM algorithm for estimating means, standard deviations, and the correlation of the bivariate normal distribution that gave the data in Figure 9.3.*

Step	μ_1	μ_2	σ_1	σ_2	ρ
1	1.86	.463	1.08	.738	.162
2	1.86	.707	1.08	.622	.394
3	1.86	.843	1.08	.611	.574
4	1.86	.923	1.08	.636	.679
5	1.86	.971	1.08	.667	.736
6	1.86	1.002	1.08	.694	.769
7	1.86	1.023	1.08	.716	.789
8	1.86	1.036	1.08	.731	.801
9	1.86	1.045	1.08	.743	.808
10	1.86	1.051	1.08	.751	.813
11	1.86	1.055	1.08	.756	.816
12	1.86	1.058	1.08	.760	.819
13	1.86	1.060	1.08	.763	.820
14	1.86	1.061	1.08	.765	.821
15	1.86	1.062	1.08	.766	.822
16	1.86	1.063	1.08	.767	.822
17	1.86	1.064	1.08	.768	.823
18	1.86	1.064	1.08	.768	.823
19	1.86	1.064	1.08	.769	.823
20	1.86	1.064	1.08	.769	.823

whose notional likelihood $\prod_1^K f_\theta(\boldsymbol{x}^{*k})$ yields MLE $\hat{\theta}^*$. It then turns out that $\hat{\theta}^*$ goes to $\hat{\theta}(\boldsymbol{o})$ as K goes to infinity. In other words, maximum likelihood estimation is *self-consistent*: generating artificial data from the MLE density $f_{\hat{\theta}(\boldsymbol{o})}(\boldsymbol{u}|\boldsymbol{o})$ doesn't change the MLE. Moreover, any value $\hat{\theta}^{(0)}$ not equal to the MLE $\hat{\theta}(\boldsymbol{o})$ cannot be self-consistent: carrying through (9.45)–(9.46) using $f_{\hat{\theta}^{(0)}}(\boldsymbol{u}|\boldsymbol{o})$ leads to hypothetical MLE $\hat{\theta}^{(1)}$ having $f_{\hat{\theta}^{(1)}}(\boldsymbol{o}) > f_{\hat{\theta}^{(0)}}(\boldsymbol{o})$, etc., a more general version of the EM algorithm.[12]

Modern technology allows social scientists to collect huge data sets, perhaps hundreds of responses for each of thousands or even millions of individuals. Inevitably, some entries of the individual responses will be missing. *Imputation* amounts to employing some version of the fake-data principle to fill in the missing values. Imputation's goal goes beyond find-

[12] Simulation (9.45) is unnecessary in exponential families, where at each stage $data^*$ can be replaced by $(\boldsymbol{o}, E^{(j)}(\boldsymbol{u}|\boldsymbol{o}))$, with $E^{(j)}$ indicating expectation with respect to $\hat{\theta}^{(j)}$, as in (9.43).

ing the MLE, to the creation of graphs, confidence intervals, histograms, and more, using only convenient, standard complete-data methods.

Finally, returning to survival analysis, the Kaplan–Meier estimate (9.17)
†11 is itself self-consistent.† Consider the **Arm_A** censored observation 74+ in Table 9.2. We know that that patient's survival time exceeded 74. Suppose we distribute his probability mass (1/51 of the **Arm_A** sample) to the right, in accordance with the conditional distribution for $x > 74$ defined by the **Arm_A** Kaplan–Meier survival curve. It turns out that redistributing all the censored cases does not change the original Kaplan–Meier survival curve; Kaplan–Meier is self-consistent, leading to its identification as the "nonparametric MLE" of a survival function.

9.6 Notes and Details

The progression from life tables, Kaplan–Meier curves, and the log-rank test to proportional hazards regression was modest in its computational demands, until the final step. Kaplan–Meier curves lie within the capabilities of mechanical calculators. Not so for proportional hazards, which is emphatically a child of the computer age. As the algorithms grew more intricate, their inferential justification deepened in scope and sophistication. This is a pattern we also saw in Chapter 8, in the progression from bioassay to logistic regression to generalized linear models, and will reappear as we move from the jackknife to the bootstrap in Chapter 10.

Censoring is not the same as truncation. For the truncated galaxy data of Section 8.3, we learn of the existence of a galaxy only if it falls into the observation region (8.38). The censored individuals in Table 9.2 are known to exist, but with imperfect knowledge of their lifetimes. There is a version of the Kaplan–Meier curve applying to truncated data, which was developed in the astronomy literature by Lynden-Bell (1971).

The methods of this chapter apply to data that is left-truncated as well as right-censored. In a survival time study of a new HIV drug, for instance, subject i might not enter the study until some time τ_i after his or her initial diagnosis, in which case t_i would be left-truncated at τ_i, as well as possibly later right-censored. This only modifies the composition of the various risk sets. However, other missing-data situations, e.g., left- *and* right-censoring, require more elaborate, less elegant, treatments.

†1 [p. 133] *Formula* (9.10). Let the interval $[t_0, t_1]$ be partitioned into a large number of subintervals of length dt, with t_k the midpoint of subinterval k.

As in (9.4), using (9.9),

$$\begin{aligned}\Pr\{T \geq t_1 | T \geq t_0\} &\doteq \prod (1 - h(t_i)\, dt) \\ &= \exp\left\{\sum \log(1 - h(t_i)\, dt)\right\} \\ &\doteq \exp\left\{-\sum h(t_i)\, dt\right\},\end{aligned} \tag{9.47}$$

which, as $dt \to 0$, goes to (9.10).

†2 [p. 136] *Kaplan–Meier estimate.* In the life table formula (9.6) (with $k = 1$), let the time unit be small enough to make each bin contain at most one value $t_{(k)}$ (9.16). Then at $t_{(k)}$,

$$\hat{h}_{(k)} = \frac{d_{(k)}}{n - k + 1}, \tag{9.48}$$

giving expression (9.17).

†3 [p. 137] *Greenwood's formula* (9.18). In the life table formulation of Section 9.1, (9.6) gives

$$\log \hat{S}_j = \sum_1^j \log\left(1 - \hat{h}_k\right). \tag{9.49}$$

From $\hat{h}_k \overset{\text{ind}}{\sim} \text{Bi}(n_k, h_k)$ we get

$$\begin{aligned}\text{var}\left\{\log \hat{S}_j\right\} &= \sum_1^j \text{var}\left\{\log\left(1 - \hat{h}_k\right)\right\} \doteq \sum_1^j \frac{\text{var}\,\hat{h}_k}{(1 - h_k)^2} \\ &= \sum_1^j \frac{h_k}{1 - h_k} \frac{1}{n_k},\end{aligned} \tag{9.50}$$

where we have used the delta-method approximation $\text{var}\{\log X\} \doteq \text{var}\{X\}/E\{X\}^2$. Plugging in $\hat{h}_k = y_k/n_k$ yields

$$\text{var}\left\{\log \hat{S}_j\right\} \doteq \sum_1^j \frac{y_k}{n_k(n_k - y_k)}. \tag{9.51}$$

Then the inverse approximation $\text{var}\{X\} = E\{X\}^2 \text{var}\{\log X\}$ gives Greenwood's formula (9.18).

The censored data situation of Section 9.2 does not enjoy independence between the $\hat{h}_k$ values. However, successive conditional independence, given the n_k values, is enough to verify the result, as in the partial likelihood calculations below. *Note*: the confidence intervals in Figure 9.1 were obtained

by exponentiating the intervals,

$$\log \hat{S}_j \pm 1.96 \left[\operatorname{var} \left\{ \log \hat{S}_j \right\} \right]^{1/2}. \tag{9.52}$$

†4 [p. 138] *Parametric life tables analysis.* Figure 9.2 and the analysis behind it is developed in Efron (1988), where it is called "partial logistic regression" in analogy with partial likelihood.

†5 [p. 139] *The log-rank test.* This chapter featured an all-star cast, including four of the most referenced papers of the post-war era: Kaplan and Meier (1958), Cox (1972) on proportional hazards, Dempster *et al.* (1977) codifying and naming the EM algorithm, and Mantel and Haenszel (1959) on the log-rank test. (Cox (1958) gives a careful, and early, analysis of the Mantel–Haenszel idea.) The not very helpful name "log-rank" does at least remind us that the test depends only on the ranks of the survival times, and will give the same result if all the observed survival times t_i are monotonically transformed, say to $\exp(t_i)$ or $t_i^{1/2}$. It is often referred to as the Mantel–Haenszel or Cochran–Mantel–Haenszel test in older literature. Kaplan–Meier and proportional hazards are also rank-based procedures.

†6 [p. 141] *Hypergeometric distribution.* Hypergeometric calculations, as for Table 9.5, are often stated as follows: n marbles are placed in an urn, n_A labeled A and n_B labeled B; n_d marbles are drawn out at random; y is the number of these labeled A. Elementary (but not simple) calculations then produce the conditional distribution of y given the table's marginals n_A, n_B, n, n_d, and n_s,

$$\Pr\{y|\text{marginals}\} = \binom{n_A}{y}\binom{n_B}{n_d - y}\Big/\binom{n}{n_d} \tag{9.53}$$

for

$$\max(n_A - n_s, 0) \le y \le \min(n_d, n_A),$$

and expressions (9.24) for the mean and variance. If n_A and n_B go to infinity such that $n_A/n \to p_A$ and $n_B/n \to 1 - p_A$, then $V \to n_d p_A(1 - p_A)$, the variance of $y \sim \text{Bi}(n_d, p_A)$.

†7 [p. 141] *Log-rank statistic* Z (9.25). Why is $(\sum_1^N V_i)^{1/2}$ the correct denominator for Z? Let $u_i = y_i - E_i$ in (9.30), so Z's numerator is $\sum_1^N u_i$, with

$$u_i | \boldsymbol{D}_i \sim (0, V_i) \tag{9.54}$$

under the null hypothesis of equal hazard rates. This implies that, unconditionally, $E\{u_i\} = 0$. For $j < i$, u_j is a function of $\boldsymbol{D}_i$ (since y_j and

E_j are), so $E\{u_j u_i | \boldsymbol{D}_i\} = 0$, and, again unconditionally, $E\{u_j u_i\} = 0$. Therefore, assuming equal hazard rates,

$$E\left(\sum_1^N u_i\right)^2 = E\left(\sum_1^N u_i^2\right) = \sum_1^N \text{var}\{u_i\} \doteq \sum_1^N V_i. \tag{9.55}$$

The last approximation, replacing unconditional variances $\text{var}\{u_i\}$ with conditional variances V_i, is justified in Crowley (1974), as is the asymptotic normality (9.29).

†8 [p. 145] *Lemma* (9.37). For $i \in \mathcal{R}_j$, the probability p_i that death occurs in the infinitesimal interval $(T_{(j)}, T_{(j)} + dT)$ is $h_i(T_{(j)})\, dT$, so

$$p_i = h_0(T_{(j)}) e^{c_i'\beta}\, dT, \tag{9.56}$$

and the probability of event A_i that individual i dies while the others don't is

$$P_i = p_i \prod_{k \in \mathcal{R}_j - i} (1 - p_k). \tag{9.57}$$

But the A_i are disjoint events, so, given that $\cup A_i$ has occurred, the probability that it is individual i who died is

$$P_i \Big/ \sum_{\mathcal{R}_j} P_j \doteq e^{c_i\beta} \Big/ \sum_{k \in \mathcal{R}_j} e^{c_k\beta}, \tag{9.58}$$

this becoming exactly (9.37) as $dT \to 0$.

†9 [p. 145] *Partial likelihood* (9.40). Cox (1975) introduced partial likelihood as inferential justification for the proportional hazards model, which had been questioned in the literature. Let $\boldsymbol{D}_j$ indicate all the observable information available just before time $T_{(j)}$(9.35), including all the death or loss times for individuals having $t_i < T_{(j)}$. (Notice that $\boldsymbol{D}_j$ determines the risk set $\mathcal{R}_j$.) By successive conditioning we write the full likelihood $f_\theta(\text{data})$ as

$$\begin{aligned} f_\theta(\text{data}) &= f_\theta(\boldsymbol{D}_1) f_\theta(i_1|\mathcal{R}_1) f_\theta(\boldsymbol{D}_2|\boldsymbol{D}_1) f_\theta(i_2|\mathcal{R}_2) \ldots \\ &= \prod_{j=1}^J f_\theta(\boldsymbol{D}_j|\boldsymbol{D}_{j-1}) \prod_{j=1}^J f_\theta(i_j|\mathcal{R}_j). \end{aligned} \tag{9.59}$$

Letting $\theta = (\alpha, \beta)$, where α is a nuisance parameter vector having to do

with the occurrence and timing of events between observed deaths,

$$f_{\alpha,\beta}(\text{data}) = \left[\prod_{j=1}^{J} f_{\alpha,\beta}(\boldsymbol{D}_j|\boldsymbol{D}_{j-1})\right] L(\beta), \tag{9.60}$$

where $L(\beta)$ is the partial likelihood (9.38).

The proportional hazards model simply ignores the bracketed factor in (9.60); $l(\beta) = \log L(\beta)$ is treated as a genuine likelihood, maximized to give $\hat{\beta}$, and assigned covariance matrix $(-\ddot{l}(\hat{\beta}))^{-1}$ as in Section 4.3. Efron (1977) shows this tactic is highly efficient for the estimation of β.

†10 [p. 148] *Fake-data principle.* For any two values of the parameters θ_1 and θ_2 define

$$l_{\theta_1}(\theta_2) = \int [\log f_{\theta_2}(\boldsymbol{o},\boldsymbol{u})]\, f_{\theta_1}(\boldsymbol{u}|\boldsymbol{o})\, d\boldsymbol{u}, \tag{9.61}$$

this being the limit as $K \to \infty$ of

$$l_{\theta_1}(\theta_2) = \lim_{K\to\infty} \frac{1}{K} \sum_{k=1}^{K} \log f_{\theta_2}(\boldsymbol{o},\boldsymbol{u}^{*k}), \tag{9.62}$$

the fake-data log likelihood (9.46) under θ_2, if θ_1 were the true value of θ. Using $f_\theta(\boldsymbol{o},\boldsymbol{u}) = f_\theta(\boldsymbol{o})\, f_\theta(\boldsymbol{u}|\boldsymbol{o})$, definition (9.61) gives

$$\begin{aligned} l_{\theta_1}(\theta_2) - l_{\theta_1}(\theta_1) &= \log\left(\frac{f_{\theta_2}(\boldsymbol{o})}{f_{\theta_1}(\boldsymbol{o})}\right) + \int \log\left(\frac{f_{\theta_2}(\boldsymbol{u}|\boldsymbol{o})}{f_{\theta_1}(\boldsymbol{u}|\boldsymbol{o})}\right) f_{\theta_1}(\boldsymbol{u}|\boldsymbol{o}) \\ &= \log\left(\frac{f_{\theta_2}(\boldsymbol{o})}{f_{\theta_1}(\boldsymbol{o})}\right) - \frac{1}{2} D\left(f_{\theta_1}(\boldsymbol{u}|\boldsymbol{o}),\, f_{\theta_2}(\boldsymbol{u}|\boldsymbol{o})\right), \end{aligned} \tag{9.63}$$

with D the deviance (8.31), which is always positive unless $\boldsymbol{u}|\boldsymbol{o}$ has the same distribution under θ_1 and θ_2, which we will assume doesn't happen.

Suppose we begin the EM algorithm at $\theta = \theta_1$ and find the value θ_2 maximizing $l_{\theta_1}(\theta)$. Then $l_{\theta_1}(\theta_2) > l_{\theta_1}(\theta_1)$ and $D > 0$ implies $f_{\theta_2}(\boldsymbol{o}) > f_{\theta_1}(\boldsymbol{o})$ in (9.63); that is, we have increased the likelihood of the observed data. Now take $\theta_1 = \hat{\theta} = \arg\max_\theta f_\theta(\boldsymbol{o})$. Then the right side of (9.63) is negative, implying $l_{\hat{\theta}}(\hat{\theta}) > l_{\hat{\theta}}(\theta_2)$ for any θ_2 not equaling $\theta_1 = \hat{\theta}$. Putting this together,[13] successively computing $\theta_1, \theta_2, \theta_3, \ldots$ by fake-data MLE calculations increases $f_\theta(\boldsymbol{o})$ at every step, and the only stable point of the algorithm is at $\theta = \hat{\theta}(\boldsymbol{o})$.

†11 [p. 150] *Kaplan–Meier self-consistency.* This property was verified in Efron (1967), where the name was coined.

[13] Generating the fake data is equivalent to the E step of the algorithm, the M step being the maximization of $l_{\theta_j}(\theta)$.