Chapter 9

Survival Analysis and the EM Algorithm

Survival analysis had its roots in governmental and actuarial statistics, spanning centuries of use in assessing life expectencies, 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 estimates, 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 “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 age 59. The column labeled “Surv” 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 age 30 (the beginning of the table) surviving past age 59, etc. Surv is calculated according to an ancient but ingenious algorithm.

Let \(X\) represent a typical lifetime, so

\[ f_i = \Pr\{ X = i \} \]  

\(^1\)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 males, per Social Security Administration data.
Table 9.1: Insurance company life table; at each age, \( n = \# \) policy holders, \( y = \# \) deaths, \( \hat{h} = \) hazard rate \( y/n \), \( \text{Surv} = \) Survival probability estimate (9.6).

<table>
<thead>
<tr>
<th>age</th>
<th>( n )</th>
<th>( y )</th>
<th>( \hat{h} )</th>
<th>( \text{Surv} )</th>
<th>age</th>
<th>( n )</th>
<th>( y )</th>
<th>( \hat{h} )</th>
<th>( \text{Surv} )</th>
<th>age</th>
<th>( n )</th>
<th>( y )</th>
<th>( \hat{h} )</th>
<th>( \text{Surv} )</th>
</tr>
</thead>
<tbody>
<tr>
<td>30</td>
<td>116</td>
<td>0</td>
<td>.000</td>
<td>1.000</td>
<td>50</td>
<td>225</td>
<td>2</td>
<td>.009</td>
<td>.948</td>
<td>70</td>
<td>100</td>
<td>3</td>
<td>.030</td>
<td>.723</td>
</tr>
<tr>
<td>31</td>
<td>44</td>
<td>0</td>
<td>.000</td>
<td>1.000</td>
<td>51</td>
<td>346</td>
<td>1</td>
<td>.003</td>
<td>.945</td>
<td>71</td>
<td>155</td>
<td>4</td>
<td>.026</td>
<td>.704</td>
</tr>
<tr>
<td>32</td>
<td>95</td>
<td>0</td>
<td>.000</td>
<td>1.000</td>
<td>52</td>
<td>370</td>
<td>2</td>
<td>.005</td>
<td>.940</td>
<td>72</td>
<td>92</td>
<td>1</td>
<td>.011</td>
<td>.696</td>
</tr>
<tr>
<td>33</td>
<td>97</td>
<td>0</td>
<td>.000</td>
<td>1.000</td>
<td>53</td>
<td>568</td>
<td>4</td>
<td>.007</td>
<td>.933</td>
<td>73</td>
<td>90</td>
<td>1</td>
<td>.011</td>
<td>.689</td>
</tr>
<tr>
<td>34</td>
<td>120</td>
<td>0</td>
<td>.000</td>
<td>1.000</td>
<td>54</td>
<td>1081</td>
<td>8</td>
<td>.007</td>
<td>.927</td>
<td>74</td>
<td>110</td>
<td>2</td>
<td>.018</td>
<td>.676</td>
</tr>
<tr>
<td>35</td>
<td>71</td>
<td>0</td>
<td>.014</td>
<td>.986</td>
<td>55</td>
<td>1042</td>
<td>4</td>
<td>.002</td>
<td>.925</td>
<td>75</td>
<td>108</td>
<td>5</td>
<td>.041</td>
<td>.648</td>
</tr>
<tr>
<td>36</td>
<td>125</td>
<td>0</td>
<td>.000</td>
<td>1.000</td>
<td>56</td>
<td>1094</td>
<td>10</td>
<td>.009</td>
<td>.916</td>
<td>76</td>
<td>138</td>
<td>8</td>
<td>.058</td>
<td>.611</td>
</tr>
<tr>
<td>37</td>
<td>122</td>
<td>0</td>
<td>.000</td>
<td>1.000</td>
<td>57</td>
<td>597</td>
<td>4</td>
<td>.007</td>
<td>.910</td>
<td>77</td>
<td>46</td>
<td>0</td>
<td>.000</td>
<td>.611</td>
</tr>
<tr>
<td>38</td>
<td>82</td>
<td>0</td>
<td>.000</td>
<td>1.000</td>
<td>58</td>
<td>359</td>
<td>1</td>
<td>.003</td>
<td>.908</td>
<td>78</td>
<td>75</td>
<td>4</td>
<td>.053</td>
<td>.578</td>
</tr>
<tr>
<td>39</td>
<td>113</td>
<td>0</td>
<td>.000</td>
<td>1.000</td>
<td>59</td>
<td>312</td>
<td>5</td>
<td>.016</td>
<td>.893</td>
<td>79</td>
<td>69</td>
<td>6</td>
<td>.087</td>
<td>.528</td>
</tr>
<tr>
<td>40</td>
<td>79</td>
<td>0</td>
<td>.000</td>
<td>1.000</td>
<td>60</td>
<td>231</td>
<td>1</td>
<td>.004</td>
<td>.889</td>
<td>80</td>
<td>95</td>
<td>4</td>
<td>.042</td>
<td>.506</td>
</tr>
<tr>
<td>41</td>
<td>90</td>
<td>0</td>
<td>.000</td>
<td>1.000</td>
<td>61</td>
<td>245</td>
<td>5</td>
<td>.020</td>
<td>.871</td>
<td>81</td>
<td>124</td>
<td>6</td>
<td>.048</td>
<td>.481</td>
</tr>
<tr>
<td>42</td>
<td>154</td>
<td>0</td>
<td>.000</td>
<td>1.000</td>
<td>62</td>
<td>196</td>
<td>5</td>
<td>.026</td>
<td>.849</td>
<td>82</td>
<td>67</td>
<td>7</td>
<td>.104</td>
<td>.431</td>
</tr>
<tr>
<td>43</td>
<td>103</td>
<td>0</td>
<td>.000</td>
<td>1.000</td>
<td>63</td>
<td>180</td>
<td>4</td>
<td>.022</td>
<td>.830</td>
<td>83</td>
<td>112</td>
<td>12</td>
<td>.107</td>
<td>.385</td>
</tr>
<tr>
<td>44</td>
<td>144</td>
<td>0</td>
<td>.000</td>
<td>1.000</td>
<td>64</td>
<td>170</td>
<td>2</td>
<td>.012</td>
<td>.820</td>
<td>84</td>
<td>113</td>
<td>8</td>
<td>.071</td>
<td>.358</td>
</tr>
<tr>
<td>45</td>
<td>192</td>
<td>2</td>
<td>.010</td>
<td>.976</td>
<td>65</td>
<td>114</td>
<td>0</td>
<td>.000</td>
<td>.820</td>
<td>85</td>
<td>116</td>
<td>12</td>
<td>.103</td>
<td>.321</td>
</tr>
<tr>
<td>46</td>
<td>153</td>
<td>1</td>
<td>.007</td>
<td>.969</td>
<td>66</td>
<td>185</td>
<td>5</td>
<td>.027</td>
<td>.798</td>
<td>86</td>
<td>124</td>
<td>17</td>
<td>.137</td>
<td>.277</td>
</tr>
<tr>
<td>47</td>
<td>179</td>
<td>1</td>
<td>.006</td>
<td>.964</td>
<td>67</td>
<td>127</td>
<td>2</td>
<td>.016</td>
<td>.785</td>
<td>87</td>
<td>110</td>
<td>21</td>
<td>.191</td>
<td>.224</td>
</tr>
<tr>
<td>48</td>
<td>210</td>
<td>0</td>
<td>.000</td>
<td>1.000</td>
<td>68</td>
<td>127</td>
<td>5</td>
<td>.039</td>
<td>.755</td>
<td>88</td>
<td>63</td>
<td>9</td>
<td>.143</td>
<td>.192</td>
</tr>
<tr>
<td>49</td>
<td>259</td>
<td>2</td>
<td>.008</td>
<td>.956</td>
<td>69</td>
<td>158</td>
<td>2</td>
<td>.013</td>
<td>.745</td>
<td>89</td>
<td>79</td>
<td>10</td>
<td>.127</td>
<td>.168</td>
</tr>
</tbody>
</table>

is the probability of dying at age \( i \), and

\[
S_i = \sum_{j \geq i} f_j = \Pr\{X \geq i\} \quad (9.2)
\]

is the probability of surviving past age \( i - 1 \). The hazard rate at age \( i \) is by definition

\[
h_i = f_i/S_i = \Pr\{X = i|X \geq i\}, \quad (9.3)
\]

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,

\[
S_{ij} = \prod_{k=i}^{j} (1 - h_k) = \Pr\{X > j|X \geq i\}; \quad (9.4)
\]

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 \( S_{1,i-1} \).

\( \text{Surv} \) 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,

\[
\hat{h}_i = y_i/n_i, \quad (9.5)
\]
and then setting

\[ \hat{S}_{30,j} = \prod_{k=30}^{j} (1 - \hat{h}_k). \] (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,

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

The hazard rate

\[ h(t) = f(t)/S(t) \] (9.8)
satisfies

\[ h(t) \, dt = \Pr\{T \in (t, t + dt) \mid T \geq t\} \] (9.9)

for \( dt \to 0 \), in analogy with (9.3). The analog of (9.4) is\(^†\)

\[ \Pr\{T \geq t_1 \mid T \geq t_0\} = \exp\left\{ - \int_{t_0}^{t_1} h(x) \, dx \right\} \] (9.10)

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

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

(As before, † indicates material discussed further in the chapter end notes.)

A one-sided exponential density

\[ f(t) = \frac{1}{c} e^{-t/c} \quad \text{for } t \geq 0 \] (9.12)

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

\[ h(t) = 1/c. \] (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.

---

\(^3\)Of course the estimates can go badly wrong if the hazard rates change over time.
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. There are patients “lost to followup”, mostly because the NCOG experiment ended with some of the patients still alive.

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

<table>
<thead>
<tr>
<th>Arm A: Chemotherapy</th>
<th>Arm B: Chemotherapy+Radiation</th>
</tr>
</thead>
<tbody>
<tr>
<td>7 34 42 63 64 74+ 83 84 91 108 112</td>
<td>37 84 92 94 110 112 119 127 130 133 140</td>
</tr>
<tr>
<td>129 133 139 140 140 146 149 154 157 160</td>
<td>146 155 159 169+ 173 179 194 195 209 240 281</td>
</tr>
<tr>
<td>160 165 173 176 185+ 218 225 241 248 273 277</td>
<td>319 339 469 519 528+ 547+ 613+ 633 725 759+</td>
</tr>
<tr>
<td>279+ 297 319+ 405 417 420 440 523 523+ 583 594</td>
<td>817 1092+ 1245+ 1331+ 1557 1642+ 1771+ 1776 1897+ 2023+ 2146+</td>
</tr>
<tr>
<td>1101 1116+ 1146 1226+ 1349+ 1412+ 1417</td>
<td>2297+ 2897+</td>
</tr>
</tbody>
</table>

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 if 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 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”

---

\(^4\)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 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 deaths, \( l \) = lost to followup, \( h \) = hazard rate \( y/n \); Surv = life table survival estimate.

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

at the beginning of month 5, etc.

“Surv” 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 \( x_i \) for censored data problems are of the form

\[
x_i = (t_i, d_i),
\]

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

\[
d_i = \begin{cases} 
1 & \text{if death observed} \\
0 & \text{if death not observed} 
\end{cases}
\]

(9.15)
(so $d_i = 0$ corresponds to a $+$ in Table 9.2). Let

$$t_{(1)} < t_{(2)} < t_{(3)} < \cdots < t_{(n)}$$  \hspace{1cm} (9.16)$$
de note the ordered survival times,\footnote{Assuming no ties among the survival times, which is convenient but not crucial for what follows.} censored or not, with corresponding indicator $d_{(k)}$ for $t_{(k)}$. The Kaplan–Meier estimate for survival probability $S(j) = \Pr\{X > t(j)\}$ is then\footnote{Assuming no ties among the survival times, which is convenient but not crucial for what follows.}

$$\hat{S}(j) = \prod_{k \leq j} \left( \frac{n - k}{n - k + 1} \right)^{d_{(k)}}. \hspace{1cm} (9.17)$$

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

---

**Figure 9.1:** NCOG Kaplan–Meier survival curves; lower Arm A (chemo only); upper Arm B (chemo+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 to 182 days for Arm A. 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\footnote{Assuming no ties among the survival times, which is convenient but not crucial for what follows.} for its standard deviation (now back in life table notation),

$$\text{sd} \left( \hat{S}(j) \right) = \hat{S}(j) \left[ \sum_{k \leq j} \frac{y_k}{n_k(n_k - y_k)} \right]^{1/2}. \hspace{1cm} (9.18)$$
9.2. CENSORED DATA AND THE KAPLAN–MEIER ESTIMATE

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.\(^\dagger\) Reverting to the life table form of Table 9.3, we assume that the death counts \( y_k \) are independent binomials,

\[
y_k \overset{\text{ind}}{\sim} \text{Bi}(n_k, h_k),
\]

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

\[
\lambda = X\alpha,
\]

as in (8.22). For instance a cubic regression would have \( x_k = (1, k, k^2, k^3) \) for the \( k \)th row of \( X \), with \( X \ 47 \times 4 \) for Table 9.3.

\(\dagger\) Reprinted with permission.
The parametric hazard rate estimates in Figure 9.2 were instead based on a “cubic-linear spline”,

\[ x_k = (1, k, (k - 11)^2, (k - 11)^3), \]  

(9.21)

where \((k - 11)^-\) equals \(k - 11\) for \(k \leq 11\), and 0 for \(k \geq 11\). The vector \(\lambda = X\alpha\) 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

\[ \hat{h}_k = \frac{1}{1 + e^{-x_k\hat{\alpha}}} \]  

(9.22)
as in (8.8). The dotted curve in Figure 9.2 traces \(\hat{h}_k\) for Arm A, while the undotted 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 as 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 are suitable for censored data. The log rank test\(^7\) 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, 50 in Arm A, with 1 and 2 deaths during the month respectively. (Losses 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.)

*\(^6\)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 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 first 6 months of NCOG study. For example, at the beginning of the 6th month after treatment, there were 33 remaining Arm B patients of whom 4 died during the month, compared to 32 at risk and 7 dying in Arm A. Conditional expected number of deaths in Arm A, assuming null hypothesis of equal hazard rates in both arms, was 5.42 (9.24).

<table>
<thead>
<tr>
<th>Month</th>
<th>Arm B</th>
<th>Arm A</th>
<th>Exp #</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>at risk</td>
<td>died</td>
<td>at risk</td>
</tr>
<tr>
<td>1</td>
<td>45</td>
<td>0</td>
<td>51</td>
</tr>
<tr>
<td>2</td>
<td>45</td>
<td>1</td>
<td>50</td>
</tr>
<tr>
<td>3</td>
<td>44</td>
<td>1</td>
<td>48</td>
</tr>
<tr>
<td>4</td>
<td>43</td>
<td>5</td>
<td>42</td>
</tr>
<tr>
<td>5</td>
<td>38</td>
<td>5</td>
<td>40</td>
</tr>
<tr>
<td>6</td>
<td>33</td>
<td>4</td>
<td>32</td>
</tr>
</tbody>
</table>

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 $n_A, n_B, n_d, n_s$ 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 NCOG study. $E$ is expected number of Arm A deaths assuming null hypothesis of equal hazard rates (last column of Table 9.4).

<table>
<thead>
<tr>
<th>Died</th>
<th>Survived</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>$y = 7$</td>
</tr>
<tr>
<td></td>
<td>$E = 5.42$</td>
</tr>
<tr>
<td>B</td>
<td>4</td>
</tr>
<tr>
<td></td>
<td>$n_d = 11$</td>
</tr>
<tr>
<td></td>
<td>$n_B = 33$</td>
</tr>
<tr>
<td></td>
<td>$n = 65$</td>
</tr>
</tbody>
</table>

Consider the null hypothesis that the hazard rates (9.3) for month 6 are the same in Arm A and Arm B,

$$H_0(6) : h_{A6} = h_{B6}.$$  \hspace{1cm} (9.23)

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

$$E = n_A n_d / n$$
$$V = n_A n_B n_d n_s / [n^2 (n - 1)],$$  \hspace{1cm} (9.24)
as calculated according to the hypergeometric distribution.

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 \[ Z = \frac{\sum_{i=1}^{N} (y_i - E_i)}{\left( \sum_{i=1}^{N} V_i \right)^{1/2}}. \] (9.25)

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

\[ H_0(i) : h_{Ai} = h_{Bi}. \] (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,

\[ \sum_{i=1}^{N} y_i = 42, \quad \sum_{i=1}^{N} E_i = 32.9, \quad \sum_{i=1}^{N} V_i = 16.0, \] (9.27)

giving log rank test statistic

\[ Z = 2.27. \] (9.28)

Asymptotic calculations based on the central limit theorem suggest \[ Z \sim \mathcal{N}(0, 1) \] (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, fineses 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, 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 $Z_i^2$’s) 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 $D_i$ indicate all the data except $y_i$ available at the end of the $i$th period. For month 6 in the NCOG study, $D_6$ includes all data for months 1–5 in Table 9.4, and the marginals $n_A, n_B, n_d, n_s$ 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),

\[ y_i \mid D_i \overset{\text{ind}}{\sim} (E_i, V_i), \quad (9.30) \]

“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 end notes, justifies adding the variances.)

The purpose of all this Fisherian conditioning is to simplify the inference: the conditional distribution $y_i \mid D_i$ 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 $x_i$ are of the form

\[ x_i = (c_i, t_i, d_i), \quad (9.31) \]

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.

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, calendar date of
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.

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

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 $i$th individual (9.8) is

$$h_i(t) = h_0(t)e^{c_i\beta}.$$  \hspace{1cm} (9.32)

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

$$\theta_i = e^{c_i\beta};$$  \hspace{1cm} (9.33)

model (9.32) says that individual $i$’s hazard is a constant factor $\theta_i$ times the baseline hazard. Equivalently, from (9.11), the $i$th survival function $S_i(t)$ is a power of the baseline survival
9.4. THE PROPORTIONAL HAZARDS MODEL

function \( S_0(t) \),
\[
S_i(t) = S_0(t)^{\theta_i}. \tag{9.34}
\]
Larger values of \( \theta_i \) lead to more quickly declining survival curves, i.e., to worse survival.

Let \( J \) be the number of observed deaths, \( J = 160 \) here, occurring at times
\[
T(1) < T(2) < \cdots < T(J), \tag{9.35}
\]
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.

**Lemma.** \( \dagger \) 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
\[
\Pr \{ i_j = i | \mathcal{R}_j \} = e^{c_i \beta} / \sum_{k \in \mathcal{R}_j} e^{c_k \beta}. \tag{9.37}
\]
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 \( \beta \) in model \( (9.32) \), we multiply factors \( (9.37) \) to form the partial likelihood
\[
L(\beta) = \prod_{j=1}^J \left( e^{c_{i_j} \beta} / \sum_{k \in \mathcal{R}_j} e^{c_k \beta} \right). \tag{9.38}
\]
\( L(\beta) \) 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\( \dagger \) of \( l(\beta) = \log L(\beta) \), as in Section 4.3,
\[
\hat{\beta} \sim \left( \beta, \left[ -I(\hat{\beta}) \right]^{-1} \right). \tag{9.40}
\]
\( \dagger \)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 \( T(j) \).
Table 9.7: Proportional hazards analysis of pediatric cancer data (with age, entry and far standardized). Age is significantly negative, older children doing better; entry is very significantly negative, showing a declining hazard rate with calendar date of entry; far is 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 \( \exp(\beta) \).

<table>
<thead>
<tr>
<th></th>
<th>( \beta )</th>
<th>sd</th>
<th>z-val</th>
<th>p-val</th>
<th>( \exp(\beta) )</th>
<th>lo</th>
<th>up</th>
</tr>
</thead>
<tbody>
<tr>
<td>sex</td>
<td>-0.023</td>
<td>0.160</td>
<td>-1.42</td>
<td>0.160</td>
<td>0.98</td>
<td>0.71</td>
<td>1.34</td>
</tr>
<tr>
<td>race</td>
<td>0.282</td>
<td>0.169</td>
<td>1.669</td>
<td>0.095</td>
<td>1.33</td>
<td>0.95</td>
<td>1.85</td>
</tr>
<tr>
<td>age</td>
<td>-0.235</td>
<td>0.088</td>
<td>-2.64</td>
<td>0.008**</td>
<td>0.79</td>
<td>0.67</td>
<td>0.94</td>
</tr>
<tr>
<td>entry</td>
<td>-0.460</td>
<td>0.079</td>
<td>-5.85</td>
<td>0.000***</td>
<td>0.63</td>
<td>0.54</td>
<td>0.74</td>
</tr>
<tr>
<td>far</td>
<td>0.296</td>
<td>0.072</td>
<td>4.117</td>
<td>0.000***</td>
<td>1.34</td>
<td>1.17</td>
<td>1.55</td>
</tr>
</tbody>
</table>

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 ages doing better (i.e., the estimated regression coefficient is negative). However, the dramatic effects are from 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 (9.31)), rather than nonparametric as before; and the 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 of the log rank test. Has its convenience possibly come at too high a price? In fact it can be shown\(^{1}\) that inference based on the partial likelihood is highly efficient, assuming of course the correctness of the proportional hazards model (9.31).

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

\(^{10}\)Table 9.7 obtained using R program \texttt{coxph}.\n
\(^{1}\)
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.

![Figure 9.3](image-url) Figure 9.3: 40 points from a binomial normal distribution, the last 20 with $x[2]$ missing (circled).

A missing data situation is shown in Figure 9.3: $n = 40$ points have been independently sampled from a bivariate normal distribution (5.12), means $(\mu_1, \mu_2)$, variances $(\sigma_1^2, \sigma_2^2)$, and correlation $\rho$,

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

(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

$$\hat{\mu}_1 = \frac{\sum_{i=1}^{40} x_{1i}}{40}, \quad \hat{\mu}_2 = \frac{\sum_{i=1}^{40} x_{2i}}{40},$$

$$\hat{\sigma}_1 = \left[ \sum_{i=1}^{40} \frac{(x_{1i} - \hat{\mu}_1)^2}{40} \right]^{1/2}, \quad \hat{\sigma}_2 = \left[ \sum_{i=1}^{40} \frac{(x_{2i} - \hat{\mu}_2)^2}{40} \right]^{1/2},$$

(9.42)

$$\hat{\rho} = \left[ \sum_{i=1}^{40} \frac{(x_{1i} - \hat{\mu}_1)(x_{2i} - \hat{\mu}_2)}{40} \right] / (\hat{\sigma}_1 \hat{\sigma}_2),$$
are unavailable with 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 \( \text{data}^o \). Then:

- The standard method (9.42) is applied to \( \text{data}^o \) to produce \( \hat{\theta}^o = (\hat{\mu}^o_1, \hat{\mu}^o_2, \hat{\sigma}_1^o, \hat{\sigma}_2^o, \hat{\rho}^o) \); this is the M (“maximizing”) step.

- Each of the missing values is replaced by its conditional expectation (assuming \( \theta = \hat{\theta}^o \)) given the nonmissing data; this is the E (“expectation”) step. In our case the missing values \( x_{2i} \) are replaced by

\[
\hat{\mu}_2^{(0)} + \frac{\hat{\sigma}_2^{(0)}}{\hat{\sigma}_1^{(0)}} \left( x_{1i} - \hat{\mu}_1^{(0)} \right).
\]  

(9.43)

- The E and M steps are repeated, at the \( j \)th stage giving a new artificially complete data set \( \text{data}^{(j)} \) 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 \( \hat{\theta}^o \).
based on just the observed data \( o \); moreover, the likelihood \( f_{\hat{\theta}(j)}(o) \) 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. Let \( x = (o, u) \) represent the “complete data”, of which \( o \) is observed while \( u \) is unobserved or missing. Write the density for \( x \) as

\[
f_\theta(x) = f_\theta(o) f_\theta(u | o), \tag{9.44}
\]

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

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

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

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

\[
data^* = \{x^{*1}, x^{*2}, \ldots, x^{*K}\}, \tag{9.46}
\]

whose notional likelihood \( \prod_1^K f_\theta(x^{*k}) \) yields MLE \( \hat{\theta}^* \). It then turns out that \( \hat{\theta}^* \) goes to \( \hat{\theta}(o) \) as \( K \) goes to infinity. In other words, maximum likelihood estimation is self-consistent: generating artificial data from the MLE density \( f_{\hat{\theta}(o)}(u | o) \) doesn’t change the MLE. Moreover, any value \( \hat{\theta}^{(0)} \) not equal to the MLE \( \hat{\theta}(o) \) is not self-consistent†: carrying through (9.45)–(9.46) using \( f_{\hat{\theta}^{(0)}}(u | o) \) leads to hypothetical MLE \( \hat{\theta}^{(1)} \) having \( f_{\hat{\theta}^{(1)}}(o) > f_{\hat{\theta}^{(0)}}(o) \) etc., a more general version of the EM algorithm.\textsuperscript{11}

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

Finally, returning to survival analysis, the Kaplan–Meier estimate (9.17) 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.

\textsuperscript{11}Simulation (9.45) is unnecessary in exponential families, where at each stage \( data^* \) can be replaced by \((o, E^{(j)}(u | o)), E^{(j)} \) indicating expectation with respect to \( \hat{\theta}^{(j)} \), as in (9.43).
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 and log rank tests 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.

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, Laird and Rubin (1977) codifying and naming the EM algorithm, and Mantel and Haenszel (1959) on the log rank test. 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.

Censoring is not the same as truncation. For the truncated galaxy data of Section 8.3, we only learn of the existence of a galaxy if it falls into the observation region (8.39). 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, 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 \( \tau_i \) after his or her initial diagnosis, in which case \( t_i \) would be left-truncated at \( \tau_i \), 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.

\textit{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),

\[
\Pr\{T \geq t_1 | T \geq t_0\} = \prod (1 - h(t_i)dt) = \exp\left\{ \sum \log (1 - h(t_i)dt) \right\} = \exp\left\{ - \sum h(t_i)dt \right\}, \tag{9.47}
\]

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

\textit{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).

Greenwood’s formula (9.18) In the life table formulation of Section 9.1, (9.6) gives

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

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

$$\text{var} \left\{ \log \hat{S}_j \right\} = \sum_1^j \text{var} \left\{ \log \left(1 - \hat{h}_k\right) \right\} = \sum_1^j \frac{\hat{h}_k}{(1 - h_k)^2} \quad (9.50)$$

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

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

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,

$$\log \hat{S}_j \pm \left[1.96 \text{ var} \left\{ \log \hat{S}_j \right\} \right]^{1/2}. \quad (9.52)$$

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.

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 $n_A, n_B, n, n_d, n_s$,

$$\Pr \{ y \mid \text{marginals} \} = \binom{n_A}{y} \binom{n_B}{n_d - y} / \binom{n}{n_d} \quad (9.53)$$
for
\[ \max(n_A - n_s, 0) \leq y \leq \min(n_d, n_A), \]
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 \( V \to npA(1 - p_A) \), the variance of \( y \sim \text{Bi}(n_d, p_A) \).

Log rank statistic \( Z \) (9.25) Why is \( \left( \sum N \right)^{1/2} \) the correct denominator for \( Z \)? Let \( u_i = y_i - E_i \) in (9.30), so \( Z \)'s numerator is \( \sum N \) with
\[ u_i | D_i \sim (0, V_i) \]  
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 \( D_i \) (since \( y_j \) and \( E_j \) are), so \( E\{u_j u_i | D_i\} = 0 \), and again unconditionally, \( E\{u_j u_i\} = 0 \). Therefore, assuming equal hazard rates,
\[
E \left( \sum_{i=1}^{N} u_i \right)^2 = E \left( \sum_{i=1}^{N} u_i^2 \right) = \sum_{i=1}^{N} \text{var}\{u_i\} = \sum_{i=1}^{N} V_i. \tag{9.55}
\]

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).

Lemma (9.37) For \( \lambda \in R_j \), the probability \( p_i \) that death occurs in the infinitesimal interval \( (T_j, T_j + dT) \) is \( h_i(T_j)e^{\text{ci}\beta dT} \), so
\[
p_i = h_0(T_j)e^{\text{ci}\beta}dT, \tag{9.56}
\]
so the probability of event \( A_i \) that individual \( i \) dies while the others don’t is
\[
P_i = p_i \prod_{k \in R_j-i} (1 - p_k). \tag{9.57}
\]
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
\[
P_i \bigg/ \sum_{R_j} P_j \overset{\text{ci}\beta}{=} \sum_{k \in R_j} e^{\text{ci}\beta}, \tag{9.58}
\]
this becoming exactly (9.37) as \( dT \to 0 \).

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 \( D_j \) 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 \( D_j \) determines the risk set \( R_j \).) By successive conditioning we write the full likelihood \( f_\theta(\text{data}) \) as

\[
 f_\theta(\text{data}) = f_\theta(D_1)f_\theta(i_1|R_1)f_\theta(D_2|D_1)f_\theta(i_2|R_2) \cdots
 = \prod_{j=1}^J f_\theta(D_j|D_{j-1}) \prod_{j=1}^J f_\theta(i_j|R_j). \tag{9.59}
\]

Letting \( \theta = (\alpha, \beta) \), where \( \alpha \) is a nuisance parameter vector having to do with the occurrence and timing of events between observed deaths,

\[
 f_{\alpha,\beta}(\text{data}) = \left[ \prod_{j=1}^J f_{\alpha,\beta}(D_j|D_{j-1}) \right] L(\beta), \tag{9.60}
\]

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 \((-\hat{\beta})^{-1}\) as in Section 4.3. Efron (1977) shows this tactic is highly efficient for the estimation of \( \beta \).

**Fake data principle**  For any two values of the parameters \( \theta_1 \) and \( \theta_2 \) define

\[
 l_{\theta_1}(\theta_2) = \int \left[ \log f_{\theta_2}(o, u) \right] f_{\theta_1}(u|o) \, du, \tag{9.61}
\]

this being the limit as \( K \to \infty \) of the fake data log likelihood (9.46) under \( \theta_2 \),

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

if \( \theta_1 \) were the true value of \( \theta \). Using \( f_\theta(o, u) = f_\theta(o)f_\theta(u|o) \), definition (9.61) gives

\[
 l_{\theta_1}(\theta_2) - l_{\theta_1}(\theta_1) = \log \frac{f_{\theta_2}(o)}{f_{\theta_1}(o)} - \int \left( \log \frac{f_{\theta_2}(u|o)}{f_{\theta_1}(u|o)} \right) f_{\theta_1}(u|o) \, du
 = \log \frac{f_{\theta_2}(o)}{f_{\theta_1}(o)} - \frac{1}{2} D(f_{\theta_1}(u|o), f_{\theta_2}(u|o)), \tag{9.63}
\]

with \( D \) the deviance (8.31), always positive unless \( u|o \) has the same distribution under \( \theta_1 \) and \( \theta_2 \), which we will assume doesn’t happen.

Suppose we begin the EM algorithm at \( \theta = \theta_1 \) and find the value \( \theta_2 \) maximizing \( l_{\theta_1}(\theta) \). Then \( l_{\theta_1}(\theta_2) > l_{\theta_1}(\theta_1) \) and \( D > 0 \) implies \( f_{\theta_2}(o) > f_{\theta_1}(o) \) in (9.63); that is, we have increased the likelihood of the observed data. Now take \( \theta_1 = \hat{\theta} = \arg \max_\theta f_\theta(o) \). Then the right side of (9.63) is negative, implying \( l_{\hat{\theta}}(\theta) > l_{\hat{\theta}}(\theta_2) \) for any \( \theta_2 \) not equaling \( \theta_1 \).
Putting this together,\textsuperscript{12} successively computing $\theta_1, \theta_2, \theta_3, \ldots$ by fake data MLE calculations increases $f_\theta(o)$ at every step, and the only stable point of the algorithm is at $\theta = \hat{\theta}(o)$.

\textit{Kaplan–Meier self-consistency} This property was verified in Efron (1967), where the name was coined.

\textsuperscript{12}Generating the fake data is equivalent to the E step of the algorithm, the M step being the maximization of $l_{\theta_j}(\theta)$. 