

## Lecture 15

# Cox regression, Part II

### 15.1 Dealing with ties

Until now in this section we have been assuming that the times of events are all distinct. In situations where event times are equal, we can carry out the same computations for Cox regression, only using a modified version of the partial likelihood. Suppose  $R_i$  is the set of individuals at risk at time  $t_i$ , and  $D_i$  the set of individuals who have their event at that time. We assume that the ties are not real ties, but only the result of discreteness in the observation. Then the probability of having precisely those individuals at time  $t_i$  will depend on the order in which they actually occurred. For example, suppose there are 5 individuals at risk at the start, and two of them have their events at time  $t_1$ . If the relative risks were  $\{\rho_1, \dots, \rho_5\}$ , where  $\rho_j = e^{\beta \cdot x_j}$ , then the first term in the partial likelihood would be

$$\frac{\rho_1}{\rho_1 + \rho_2 + \rho_3 + \rho_4 + \rho_5} \cdot \frac{\rho_2}{\rho_2 + \rho_3 + \rho_4 + \rho_5} + \frac{\rho_2}{\rho_1 + \rho_2 + \rho_3 + \rho_4 + \rho_5} \cdot \frac{\rho_1}{\rho_1 + \rho_3 + \rho_4 + \rho_5}.$$

The number of terms is  $d_i!$ , so it is easy to see that this computation quickly becomes intractable.

A very good alternative — accurate and easy to compute — was proposed by B. Efron. Observe that the terms differ in the denominator merely by a small change due to the individuals lost from the risk set. If the deaths at time  $t_i$  are not a large proportion of the risk set, then we can approximate this by deducting the average of the risks that depart. In other words, in the above example, the first contribution to the partial likelihood becomes

$$\frac{\rho_1 \rho_2}{(\rho_1 + \rho_2 + \rho_3 + \rho_4 + \rho_5) \left( \frac{1}{2}(\rho_1 + \rho_2) + \rho_3 + \rho_4 + \rho_5 \right)}.$$

More generally, the partial likelihood becomes

$$L_P(\beta) = \prod_{t_i} e^{\beta \cdot \sum_{j \in D_i} x_j} \prod_{k=0}^{d_i-1} \left( \sum_{j \in R_i} e^{\beta \cdot x_j} - \frac{k}{d_i} \sum_{j \in D_i} e^{\beta \cdot x_j} \right)^{-1}.$$

We take the same approach to estimating the baseline hazard:

$$\hat{h}_0(t_i) = \sum_{k=0}^{d_i-1} \left( \sum_{j \in R_i} e^{\hat{\beta} \cdot x_j} - \frac{k}{d_i} \sum_{j \in D_i} e^{\hat{\beta} \cdot x_j} \right)^{-1}.$$

Another approach, due to Breslow, makes no correction for the progressive loss of risk in the denominator:

$$L_P^{Breslow}(\beta) = \prod_{t_i} e^{\beta \cdot \sum_{j \in D_i} x_i} \left( \sum_{j \in R_i} e^{\beta \cdot x_i} \right)^{-d_i}.$$

This approximation is always too small, and tends to shift the estimates of  $\beta$  toward 0. It is widely used as a default in software packages (SAS, not R!) for purely historical reasons.

## 15.2 Plot for PH assumption with continuous covariate

Suppose we have a continuous covariate and we wish to check the proportional hazards assumption for that covariate. We do not have natural groups of subjects with the same value of that covariate.

Provided there is sufficient data we would group the subjects in quintiles of the covariate. Then we have 5 groups and can find the Kaplan-Meier estimator for each group. As before we plot

$$\log(-\log(\widehat{S}_k(t))) \text{ v. } \log t$$

for each  $k = 1, \dots, 5$  on the same graph. There should be a roughly constant vertical separation of groups. It generally is not a wonderful method, but is better than nothing.

## 15.3 The AML example

We continue looking at the leukemia study that we started to consider in section 12.3. First, in Figure 15.1 we plot the iterated logarithm of survival against time, to test the proportional hazards assumption. The PH assumption corresponds to the two curves differing by a vertical shift. The result makes this assumption at least credible.

We code the data with covariate  $x = 0$  for the maintained group, and  $x = 1$  for the non-maintained group. Thus, the baseline hazard will correspond to the maintained group, and  $e^\beta$  will be the relative risk of the non-maintained group. From Table 12.4 we see that the Efron approximate partial likelihood is given by

$$\begin{aligned} L_P(\beta) = & \left( \frac{e^{2\beta}}{(12e^\beta + 11)(11e^\beta + 11)} \right) \left( \frac{e^{2\beta}}{(10e^\beta + 11)(9e^\beta + 11)} \right) \\ & \times \left( \frac{1}{8e^\beta + 11} \right) \left( \frac{e^\beta}{8e^\beta + 10} \right) \left( \frac{1}{7e^\beta + 10} \right) \left( \frac{1}{6e^\beta + 8} \right) \\ & \times \left( \frac{e^\beta \cdot 1}{(6e^\beta + 7)(5.5e^\beta + 6.5)} \right) \left( \frac{e^\beta}{5e^\beta + 6} \right) \left( \frac{e^\beta}{4e^\beta + 5} \right) \\ & \times \left( \frac{1}{3e^\beta + 5} \right) \left( \frac{e^\beta}{3e^\beta + 4} \right) \left( \frac{1}{2e^\beta + 4} \right) \left( \frac{e^\beta}{2e^\beta + 3} \right) \left( \frac{e^\beta}{e^\beta + 3} \right) \left( \frac{1}{2} \right) \end{aligned} \quad (1)$$

A plot of  $L_P(\beta)$  is shown in Figure 15.4.

In the one-dimensional setting it is straightforward to estimate  $\hat{\beta}$  by direct computation. We see the maximum at  $\hat{\beta} = 0.9155$  in the plot of Figure 15.4. In more complicated settings, there are good maximisation algorithms built in to the `coxph` function in the `survival` package of R. Applying this to the current problem, we obtain:

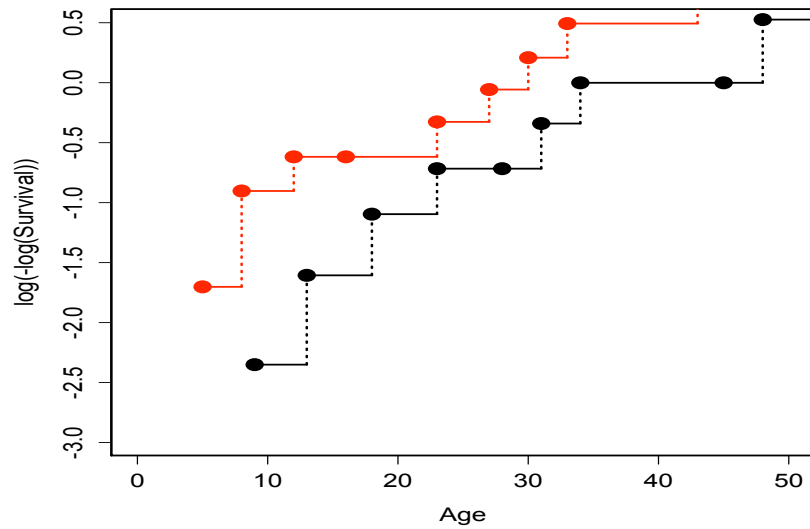


Figure 15.1: Iterated log plot of survival of two populations in AML study, to test proportional hazards assumption.

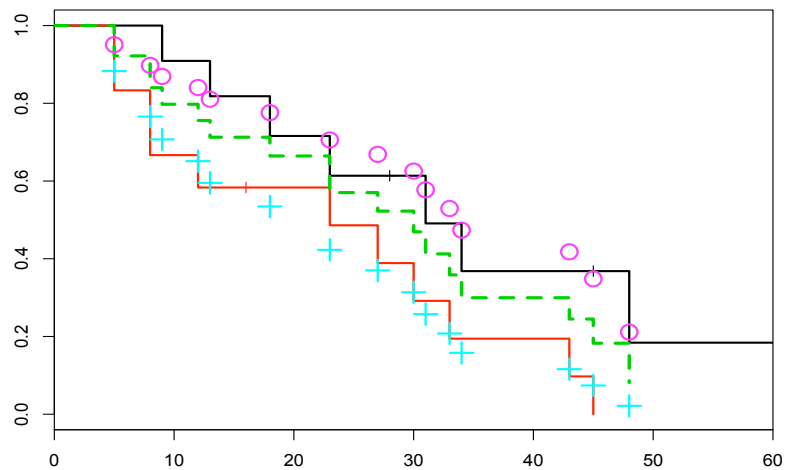


Figure 15.2: Estimated baseline hazard under the PH assumption. The purple circles show the baseline hazard; blue crosses show the baseline hazard shifted up proportionally by a multiple of  $e^{\hat{\beta}} = 2.5$ . The dashed green line shows the estimated survival rate for the mixed population (mixing the two estimates by their proportions in the initial population).

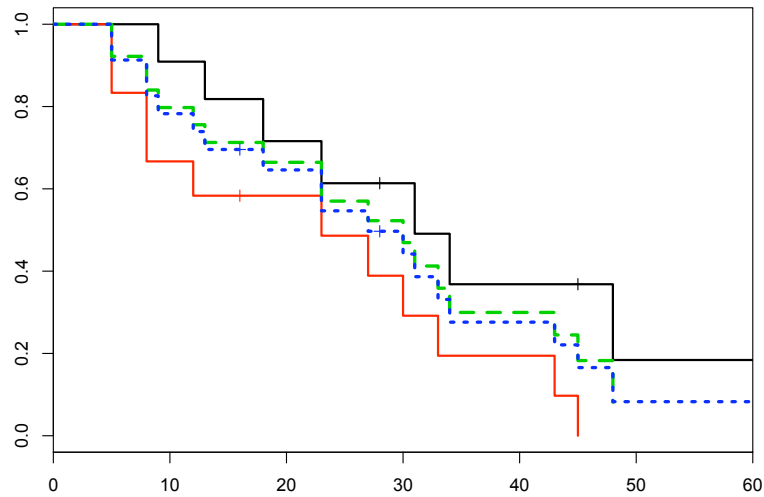


Figure 15.3: Comparing the estimated population survival under the PH assumption (green dashed line) with the estimated survival for the combined population (blue dashed line), found by applying the Nelson-Aalen estimator to the population, ignoring the covariate.

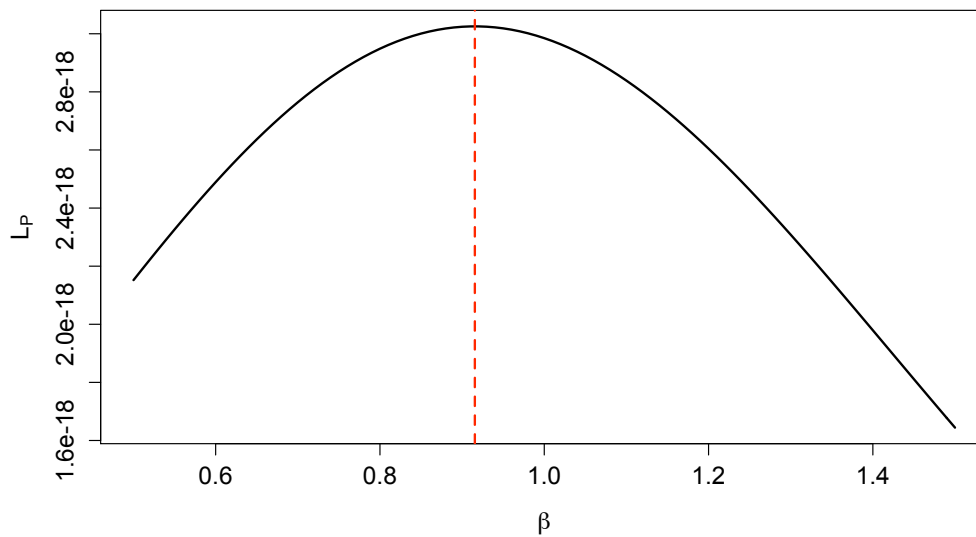


Figure 15.4: A plot of the partial likelihood from (1). Dashed line is at  $\beta = 0.9155$ .

Table 15.1: Output of the `coxph` function run on the `aml` data set.

coxph(formula = Surv(time, status) ~ x, data = aml)					
	coef	exp(coef)	se(coef)	z	p
×Nonmaintained	0.916	2.5	0.512	1.79	0.074
Likelihood ratio test=3.38 on 1 df p=0.0658 n= 23					

The  $z$  is simply the Z-statistic for testing the hypothesis that  $\beta = 0$ , so  $z = \hat{\beta}/SE(\hat{\beta})$ . We see that  $z = 1.79$  corresponds to a p-value of 0.074, so we would not reject the null hypothesis at level 0.05.

We show the estimated baseline hazard in Figure 15.2; the relevant numbers are given in Table 15.2. For example, the first hazard, corresponding to  $t_1 = 5$ , is given by

$$\hat{h}_0(5) = \frac{1}{12e^{\hat{\beta}} + 11} + \frac{1}{11e^{\hat{\beta}} + 11} = 0.050,$$

substituting in  $\hat{\beta} = 0.9155$ .

Table 15.2: Computations for the baseline hazard LME for the AML data, in the proportional hazards model, with maintained group as baseline, and relative risk  $e^{\hat{\beta}} = 2.498$ .

$t_i$	Maintenance		Non-Maintenance (control)		Baseline		
	$n_i^M$	$d_i^M$	$n_i^N$	$d_i^N$	$\hat{h}_0(t_i)$	$\hat{H}_0(t_i)$	$\tilde{S}_0(t_i)$
5	11	0	12	2	0.050	0.050	0.951
8	11	0	10	2	0.058	0.108	0.898
9	11	1	8	0	0.032	0.140	0.869
12	10	0	8	1	0.033	0.174	0.841
13	10	1	7	0	0.036	0.210	0.811
18	8	1	6	0	0.043	0.254	0.776
23	7	1	6	1	0.095	0.348	0.706
27	6	0	5	1	0.054	0.403	0.669
30	5	0	4	1	0.067	0.469	0.625
31	5	1	3	0	0.080	0.549	0.577
33	4	0	3	1	0.087	0.636	0.529
34	4	1	2	0	0.111	0.747	0.474
43	3	0	2	1	0.125	0.872	0.418
45	3	0	1	1	0.182	1.054	0.348
48	2	1	0	0	0.500	1.554	0.211