11.3 Cox Proportional Hazards Model

Definition. The **hazard function** $h(t)$ is defined as

$$h(t) = \frac{f(t)}{S(t)}.$$

It can be interpreted as the rate of dying (or experiencing an event) immediately after time t, given that the individual survived past time t. To see this, we write

$$\mathbb{P}(T < t + dt \,|\, T > t) = \frac{\mathbb{P}(t < T < t + dt)}{\mathbb{P}(T > t)} \approx \frac{f(t)dt}{S(t)} = h(t)dt.$$

Note that the survival function $S(t)$ relates to the hazard function $h(t)$ as

$$S(t) = e^{-\int_0^t h(u)du}, \; t > 0.$$

This formula is derived as a solution of the differential equation $h(t) = \dfrac{f(t)}{S(t)} = -\dfrac{S'(t)}{S(t)}$, or $\dfrac{dS(t)}{S(t)} = -h(t)dt$. Integrating both sides, we arrive at $\ln S(t) = \ln S(0) = -\displaystyle\int_0^t h(u)du$. Since $S(0) = 1$, the result follows.

Definition. Suppose that besides the event time and an indicator of censoring, data contain measurements of a set of predictors $x_1, \ldots, x_m$ that do not vary with time. Denote the event time by T and assume that it is a random variable with the hazard function $h_T(t)$. The **Cox proportional hazards model** (or simply, the **Cox model**) assumes that the hazard function has the form:

$$h_T(t, x_1, \ldots, x_m, \beta_1, \ldots, \beta_m) = h_0(t) \exp\big\{\beta_1 x_1 + \cdots + \beta_m x_m\big\}.$$

Note that in this model, the hazard function depends on time only through the **baseline hazard function** $h_0(t)$, and therefore, the ratio of hazards of two individuals doesn't depend on time, which makes the hazards are **proportional** over time.

The unknowns of this model are the baseline hazard function $h_0(t)$ and the regression coefficients $\beta_1, \ldots, \beta_m$. Instead of estimating $h_0(t)$, we introduce another formulation of the Cox model, in terms of the survival function. We write

$$S_T(t, x_1, \ldots, x_m, \beta_1, \ldots, \beta_m) = \exp\Big\{ -\int_0^t h_T(u, x_1, \ldots, x_m, \beta_1, \ldots, \beta_m)\, du \Big\}$$

$$= \exp\Big\{ - \int_0^t h_0(u) \exp\left\{\beta_1 x_1 + \cdots + \beta_m x_m\right\} du \Big\} = \Big[S_0(t)\Big]^r$$

where $S_0(t) = \exp\Big\{ - \int_0^t h_0(u)\, du \Big\}$ is the **baseline survival function**, and $r = \exp\left\{\beta_1 x_1 + \cdots + \beta_m x_m\right\}$ is the **relative risk** of an individual.

Example. Consider the data in the file "nasopharyngeal_cancer_data.csv". We fit the Cox proportional hazards model, using SAS and R.

In SAS:

```
proc phreg data=cancer_data outest=betas;
 class gender(ref="F") smoker(ref="no") therapy(ref="radio");
   model years*censored(1)= age gender smoker therapy;
     baseline out=outdata survival=Sbar;
run;

proc print data=betas;
run;
```

Analysis of Maximum Likelihood Estimates

Parameter		DF	Parameter Estimate	Standard Error	Chi-Square	Pr > ChiSq	Hazard Ratio	Label
age		1	0.04424	0.01778	6.1879	0.0129	1.045	
gender	M	1	1.25750	0.43328	8.4234	0.0037	3.517	gender M
smoker	yes	1	1.37842	0.37928	13.2084	0.0003	3.969	smoker yes
therapy	chemo	1	0.23654	0.31344	0.5695	0.4504	1.267	therapy chemo

```
proc print data=outdata;
run;
```

Obs	age	gender	smoker	therapy	years	Sbar
1	60.844155844	F	no	radio	0	1.00000
2	60.844155844	F	no	radio	0.1	0.99894
3	60.844155844	F	no	radio	0.2	0.99786
4	60.844155844	F	no	radio	1.1	0.99665
5	60.844155844	F	no	radio	1.6	0.99533
6	60.844155844	F	no	radio	1.7	0.99392
7	60.844155844	F	no	radio	1.9	0.99247
8	60.844155844	F	no	radio	2.1	0.99095
9	60.844155844	F	no	radio	2.2	0.98937
10	60.844155844	F	no	radio	2.6	0.98774
11	60.844155844	F	no	radio	2.8	0.98603
12	60.844155844	F	no	radio	3	0.98429
13	60.844155844	F	no	radio	3.5	0.98248
14	60.844155844	F	no	radio	3.6	0.97877
15	60.844155844	F	no	radio	3.8	0.97482
16	60.844155844	F	no	radio	3.9	0.97262
17	60.844155844	F	no	radio	4.2	0.97035

18	60.844155844	F	no	radio	4.7	0.96775
19	60.844155844	F	no	radio	4.8	0.96507
20	60.844155844	F	no	radio	5.1	0.96234
21	60.844155844	F	no	radio	5.2	0.95602
22	60.844155844	F	no	radio	5.3	0.94778
23	60.844155844	F	no	radio	5.5	0.94355
24	60.844155844	F	no	radio	5.6	0.93923
25	60.844155844	F	no	radio	5.7	0.93475
26	60.844155844	F	no	radio	5.8	0.92958
27	60.844155844	F	no	radio	5.9	0.92389
28	60.844155844	F	no	radio	6.1	0.91721
29	60.844155844	F	no	radio	6.3	0.91024
30	60.844155844	F	no	radio	7.2	0.90245
31	60.844155844	F	no	radio	7.3	0.89204
32	60.844155844	F	no	radio	8	0.87776
33	60.844155844	F	no	radio	8.5	0.86193
34	60.844155844	F	no	radio	8.7	0.82577
35	60.844155844	F	no	radio	9.6	0.80336
36	60.844155844	F	no	radio	10.1	0.77611
37	60.844155844	F	no	radio	10.2	0.71806
38	60.844155844	F	no	radio	10.3	0.64456
39	60.844155844	F	no	radio	10.5	0.53305

The fitted model is

$$\widehat{S}(t) = \left[\bar{S}(t)\right]^{\exp\left(0.04424(age-60.8442)+1.2575\cdot male+1.37842\cdot smoker+0.23654\cdot chemo\right)}$$

where $\bar{S}(t)$ is a step function given in the last (two) column(s) of the output.

Age at baseline, gender, and smoker are significant predictors. As age at baseline increases by one year, the hazard of dying from nasopharyngeal cancer increases by $\left(e^{0.04424}-1\right)\cdot 100\% = 4.523318\%$. The hazard for males is $100\%\cdot e^{1.2575} = 351.6619\%$ of that for females. For smokers, the hazard is $100\%\cdot e^{1.37842} = 396.8626\%$ of that for non-smokers.

Suppose we would like to predict the probability of a 5-year survival for a 60-year-old male who is a smoker and who is undergoing chemotherapy. We calculate the predicted survival function as:

$$S^0(5) = \left[\bar{S}(5)\right]^{\exp\{0.04424(60-60.8442)+1.2575+1.37842+0.23654\}} = (0.96507)^{17.03232} = 0.545759.$$

In R:

```
#fitting Cox model
gender.rel<- relevel(as.factor(cancer.data$gender), ref="F")
smoker.rel<- relevel(as.factor(cancer.data$smoker), ref="no")
therapy.rel<- relevel(as.factor(cancer.data$therapy), ref="radio")

#estimating the beta coefficients
cox.model<-coxph(Surv(years, censored==0) ~ age + gender.rel + smoker.rel
+ therapy.rel, data=cancer.data)
summary(cox.model)
```

```
                    coef exp(coef) se(coef)     z Pr(>|z|)
age              0.04482   1.04584  0.01784 2.513 0.011982
gender.relM      1.25230   3.49837  0.43260 2.895 0.003794
smoker.relyes    1.37362   3.94961  0.37929 3.622 0.000293
therapy.relchemo 0.24998   1.28400  0.31420 0.796 0.426257
```

```
#estimating the baseline survival function
base.surv<-survfit(cox.model, se.fit=FALSE)
summary(base.surv)
```

```
 time n.risk n.event survival
  0.1     77       1    0.999
  0.2     76       1    0.998
  1.1     73       1    0.997
  1.6     71       1    0.995
  1.7     70       1    0.994
  1.9     69       1    0.992
  2.1     68       1    0.991
  2.2     67       1    0.989
  2.6     66       1    0.988
  2.8     65       1    0.986
  3.0     64       1    0.984
  3.5     62       1    0.983
  3.6     61       2    0.979
  3.8     59       2    0.975
  3.9     56       1    0.973
  4.2     55       1    0.970
  4.7     52       1    0.968
```

```
 4.8     51       1    0.965
 5.1     50       1    0.962
 5.2     48       2    0.956
 5.3     46       2    0.947
 5.5     44       1    0.943
 5.6     43       1    0.939
 5.7     41       1    0.934
 5.8     38       1    0.929
 5.9     36       1    0.923
 6.1     34       1    0.917
 6.3     31       1    0.910
 7.2     28       1    0.902
 7.3     25       1    0.892
 8.0     19       1    0.877
 8.5     16       1    0.861
 8.7     14       2    0.824
 9.6     10       1    0.801
10.1      7       1    0.774
10.2      6       2    0.713
10.3      3       1    0.639
10.5      2       1    0.528
```

Note that in R, the baseline survival function is estimated at the mean values of numeric predictors (the same as in SAS). Those mean values are not given in the output. They would need to be computed separately.

mean(cancer.data$age)

```
60.84416
```

□