Lecture 6

PREDICTING SURVIVAL UNDER THE PH MODEL

The Cox PH model: $\lambda(t|\mathbf{Z}) = \lambda_0(t)\,\exp(\boldsymbol{\beta}'\mathbf{Z})$.

How do we estimate the survival probability, $S_z(t) = S(t|\mathbf{Z}) = P(T > t|\mathbf{Z})$, for an individual with covariates $\mathbf{Z}$?

For the baseline (reference) group, we have:

$$S_0(t) \;=\; e^{-\int_0^t \lambda_0(u)du} = e^{-\Lambda_0(t)}$$

For an individual with covariates $\mathbf{Z}$, we have:

$$\begin{aligned} S_z(t) &= e^{-\Lambda_z(t)} = e^{-\int_0^t \lambda_z(u)du} \\ &= e^{-\int_0^t \lambda_0(u)\exp(\beta' Z)du} \\ &= e^{-\exp(\beta' Z)\int_0^t \lambda_0(u)du} \\ &= \left[e^{-\int_0^t \lambda_0(u)du}\right]^{\exp(\beta' Z)} \\ &= [S_0(t)]^{\exp(\beta' Z)} \end{aligned}$$

Notice that in the Cox model, it is $\boldsymbol{\beta}'\mathbf{Z}$ (and not $\mathbf{Z}$) that directly determines a subject's survival distribution, i.e. two individuals with $\mathbf{Z}_1 \neq \mathbf{Z}_2$ but $\boldsymbol{\beta}'\mathbf{Z}_1 = \boldsymbol{\beta}'\mathbf{Z}_2$ have the same survival distribution. This quantity $\boldsymbol{\beta}'\mathbf{Z}$ is called the **prognostic index**.

How can we estimate the baseline survival function, $S_0(t)$?

We *could* try to use the KM estimator, but there are a few disadvantages of that approach:

- The baseline line group may not exist in data, i.e. no observation with $\mathbf{Z} = 0$.
- When it does exist, the subgroup KM would only use the survival times for observations contained in the baseline group, and not all the rest of the survival times.
- It would tend to be somewhat choppy, since it would reflect the smaller sample size of the baseline group.

Instead, we will use a baseline hazard estimator which makes use of the whole sample to get a smoother estimate. Then

$$\hat{S}_z(t) \;=\; [\hat{S}_0(t)]^{\exp(\widehat{\boldsymbol{\beta}}'\mathbf{Z})}$$

where $\widehat{\boldsymbol{\beta}}$ is the partial likelihood estimator.

We can estimate $\hat{S}_0(t)$ by one of the following approaches:

- Breslow estimator (most used)
- Kalbfleisch/Prentice estimator (read)

(1) **Breslow Estimator:**

$$\hat{S}_0(t) = \exp^{-\hat{\Lambda}_0(t)}$$

where $\hat{\Lambda}_0(t)$ is the estimated cumulative baseline hazard:

$$\hat{\Lambda}_0(t) = \sum_{j:\tau_j<t} \left(\frac{d_j}{\Sigma_{k\in\mathcal{R}(\tau_j)} \exp(\hat{\boldsymbol{\beta}}' \mathbf{Z}_k)} \right)$$

(2) **Kalbfleisch/Prentice Estimator** [read]

$$\hat{S}_0(t) = \prod_{j:\tau_j<t} \hat{\alpha}_j$$

where $\hat{\alpha}_j, j = 1, ...d$ are the MLE's obtained by assuming that $S(t|Z)$ satisfies

$$S(t|Z) = [S_0(t)]^{e^{\beta' Z}} = \left[\prod_{j:\tau_j<t} \alpha_j \right]^{e^{\beta' Z}} = \prod_{j:\tau_j<t} \alpha_j^{e^{\beta' Z}}$$

Breslow Estimator: further motivation

Heuristics: the expected number of failures in $(\tau_j, \tau_j + \Delta t)$ is

$$d_j \approx \sum_{k \in \mathcal{R}(\tau_j)} P(\tau_j < T_k < \tau_j + \Delta t | T_k > \tau_j) = \Delta t \times \sum_{k \in \mathcal{R}(\tau_j)} \lambda_0(\tau_j) \exp(\boldsymbol{\beta}' \mathbf{Z}_k)$$

Hence,

$$\Delta t \times \lambda_0(\tau_j) \approx \frac{d_j}{\Sigma_{k \in \mathcal{R}(\tau_j)} \exp(\boldsymbol{\beta}' \mathbf{Z}_k)}.$$

This leads to estimating the cumulative baseline hazard by:

$$\hat{\Lambda}_0(t) = \sum_{j:\tau_j < t} \left(\frac{d_j}{\Sigma_{k \in \mathcal{R}(\tau_j)} \exp(\hat{\boldsymbol{\beta}}' \mathbf{Z}_k)} \right)$$

Recall that for a single sample with no covariates, the **Nelson-Aalen Estimator** of the cumulative hazard is:

$$\hat{\Lambda}(t) = \sum_{j:\tau_j < t} \frac{d_j}{r_j}$$

where d_j and r_j are the number of deaths and the number at risk, respectively, at the j-th death time.

The Breslow estimator can also be derived using the profile likelihood approach (Murphy and van der Vaart, 2000).

[Read] **Kalbfleisch/Prentice Estimator: further motivation**

The idea here is a discrete time MLE (not NPMLE).
This method is analogous to the Kaplan-Meier Estimator.

If we let

$$S_0(t) = \prod_{j:\tau_j<t} (1-\lambda_j) = \prod_{j:\tau_j<t} \alpha_j$$

For someone with covariates $\mathbf{Z}$,

$$S(t;Z) = S_0(t)^{e^{\beta' Z}} = \left[\prod_{j:\tau_j<t} \alpha_j\right]^{e^{\beta' Z}} = \prod_{j:\tau_j<t} \alpha_j^{e^{\beta' Z}}$$

The likelihood contributions for discrete survival times:

- for someone censored at t: $S(t;Z)$
- for someone who fails at τ_j:

$$S(\tau_{j-1};Z) - S(\tau_j;Z) = \left[\prod_{k<j} \alpha_k\right]^{e^{\beta Z}} [1-\alpha_j^{e^{\beta Z}}]$$

From such a likelihood one can obtain the MLE $\hat{\alpha}_j$'s (K&P book p85).

In R, survfit() or survfit.coxph() gives the estimated survival curves; see documentation. Be *careful* that the software sometimes centers the covariates by giving prediction for $Z_i - \bar{Z}$ (which is typically not what we want).

When there are no covariates, the Breslow's estimate reduces to the Fleming-Harrington (Nelson-Aalen) estimate, and K/P reduces to KM.

In practice, the two estimates are very close (see Fleming and Harrington 1984, *Communications in Statistics*), although Breslow's is more commonly referred to in the survival literature.

Example

Nursinghome data: Baseline Survival Estimate

OBS	MARRIED	HEALTH	LOS	PS
1	0	0	0	1.00000
2	0	0	1	0.99253
3	0	0	2	0.98672
4	0	0	3	0.98363
5	0	0	4	0.97776
6	0	0	5	0.97012
7	0	0	6	0.96488
8	0	0	7	0.95856
9	0	0	8	0.95361
10	0	0	9	0.94793
11	0	0	10	0.94365
12	0	0	11	0.93792
13	0	0	12	0.93323
14	0	0	13	0.92706
15	0	0	14	0.92049
16	0	0	15	0.91461
17	0	0	16	0.91017
18	0	0	17	0.90534
19	0	0	18	0.90048
20	0	0	19	0.89635
21	0	0	20	0.89220
22	0	0	21	0.88727
23	0	0	22	0.88270

.
.
.

Survival Estimates by Marital and Health Status

Nursinghome data: Predicted Survival given covariates

OBS	MARRIED	HEALTH	LOS	PS
1	0	2	0	1.00000
2	0	2	1	0.98961
3	0	2	2	0.98156
4	0	2	3	0.97728
.....				
171	0	2	184	0.50104
172	0	2	185	0.49984

396	0	5	0	1.00000
397	0	5	1	0.98300
398	0	5	2	0.96988
399	0	5	3	0.96295
.....				
474	0	5	78	0.50268
475	0	5	80	0.49991

791	1	2	0	1.00000
792	1	2	1	0.98605
793	1	2	2	0.97527
794	1	2	3	0.96955
.....				
897	1	2	108	0.50114
898	1	2	109	0.49986

1186	1	5	0	1.00000
1187	1	5	1	0.97719
1188	1	5	2	0.95969
1189	1	5	3	0.95047
.....				
1233	1	5	47	0.50519
1234	1	5	48	0.49875

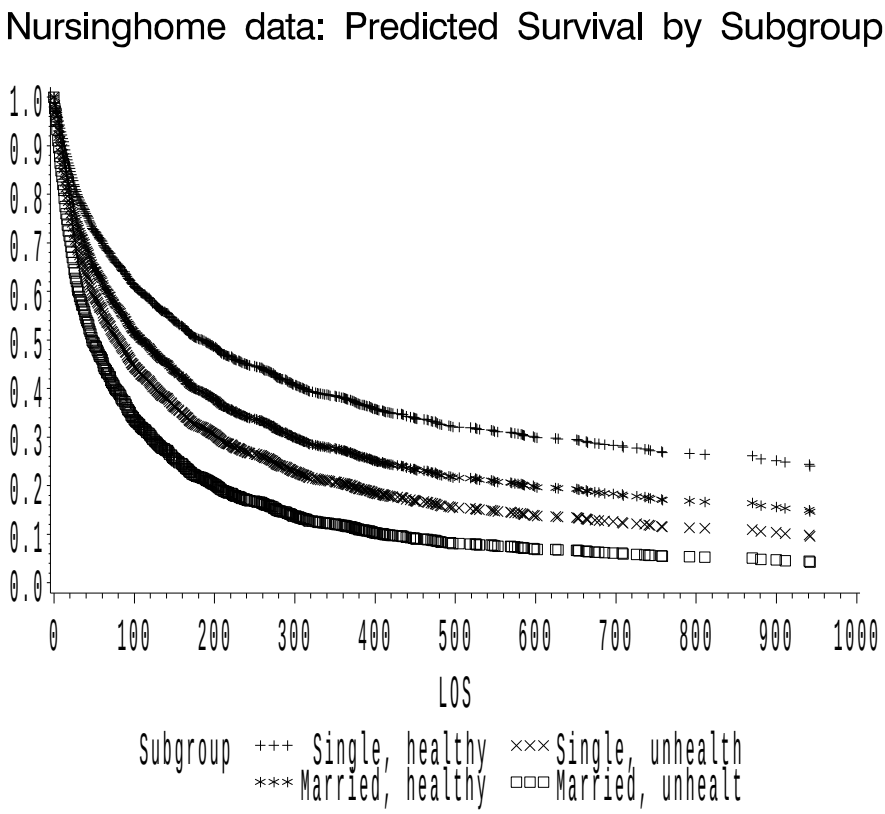

Figure 1: Nursing home estimated survival curves under the PH model

Predicted medians and P-year survival

Predicted Medians

Suppose we want to find the predicted median survival for an individual with a specified combination of covariates (e.g., a single person with health status 5).

Three possible approaches:

(1) Calculate the median from the subset of individuals with the specified covariate combination (using KM approach)

(2) Generate predicted survival curves for the combination of covariates, and obtain the medians directly

OBS	MARRIED	HEALTH	LOS	PREDSURV
171	0	2	184	0.50104
172	0	2	185	0.49984
474	0	5	78	0.50268
475	0	5	80	0.49991
897	1	2	108	0.50114
898	1	2	109	0.49986
1233	1	5	47	0.50519
1234	1	5	48	0.49875

Recall that previously we defined the median as the *smallest* value of t for which $\hat{S}(t) \leq 0.5$, so the medians from above would be 185, 80, 109, and 48 days for single healthy, single unhealthy, married healthy, and married unhealthy, respectively.

(3) Generate the predicted survival curve from the estimated baseline hazard, as follows:

We want the estimated median (M) for an individual with covariates $\mathbf{Z}$. We know

$$S(M; Z) = [S_0(M)]^{e^{\beta' Z}} = 0.5$$

Hence, M satisfies:

$$S_0(M) = [0.5]^{e^{-\beta' Z}}.$$

Eg. Suppose we want to estimate the median survival for a single unhealthy subject from the nursing home data. The reciprocal of the hazard ratio for unhealthy (health=5) is: $e^{-0.165*5} = 0.4373$, (where $\hat{\beta} = 0.165$ for health status)

So, we want M such that $S_0(M) = (0.5)^{0.4373} = 0.7385$

So the median for single unhealthy subject is the 73.8^{th} percentile of the baseline group.

OBS	MARRIED	HEALTH	LOS	PREDSURV
79	0	0	78	0.74028
80	0	0	80	0.73849
81	0	0	81	0.73670

So the estimated median would still be 80 days. Note: similar logic can be followed to estimate other quantiles besides the median.

Estimating P-year survival

Suppose we want to find the P-year survival rate for an individual with a specified combination of covariates, $\hat{S}(P; \mathbf{Z})$

For an individual with $\mathbf{Z} = 0$, the P-year survival can be obtained from the baseline survivorship function, $\hat{S}_0(P)$

For individuals with $\mathbf{Z} \neq 0$, it can be obtained as:

$$\hat{S}(P; \mathbf{Z}) = [\hat{S}_0(P)]^{e^{\widehat{\beta}' Z}}$$

Notes:

- Although we say "P-year" survival, the units of time in a particular dataset may be days, weeks, or months. The answer here will be in the same units of time as the original data.
- If $\widehat{\boldsymbol{\beta}}'\mathbf{Z}$ is positive, then the P-year survival rate for the i-th individual will be lower than for a baseline individual. (**Why is this true?**)

Estimating Subgroup Survival

(i.e. given a range of covariate values)

This is becoming useful in the era of precision medicine.

Let H be a range of covariate values, eg. age between 40-50 (while age in years was entered into the Cox model as a continuous covariate). Sometimes we want to know

$$S(t|\mathbf{Z} \in H) = P(T > t|\mathbf{Z} \in H) = ?$$

A summary of methods:

1. subgroup KM of individuals from set H: $\hat{S}_{KM}(t)$
2. 'direct adjusted survival curve'

$$\bar{S}(t) = \frac{1}{n_0} \sum_{i \in H} \hat{S}(t|\mathbf{Z}_i),$$

   where n_0 is the number of individuals in H.

3. 'average covariate method' (**Caution!**)

$$S_N(t) = \hat{S}(t|\bar{\mathbf{Z}}),$$

   where $\bar{\mathbf{Z}} = \sum_{i \in H} \mathbf{Z}_i / n_0$.

4. Xu and O'Quigley (2000) using Bayes' formula with $Z|T$: the distribution of Z given T itself is in fact the π's defined with the partial likelihood earlier.

References: Thomsen et al (1991), Xu and O'Quigley (2000)

Note: methods 1, 2 and 4 are all consistent. Method 1 only makes use of data from individuals in H, while 2 and 4 use the whole sample and the Cox model. Variance estimates are available for method 4.