9 Estimating the Underlying Survival Distribution for a Proportional Hazards Model

So far the focus has been on the regression parameters in the proportional hazards model. These parameters describe the strength of the relationship of the the covariates to survival. Since a proportional hazards model is semiparametric in the sense that the underlying baseline hazard function is left totally unspecified, these parameters do not suffice in describing the survival distribution. However, we may be interested in estimating the survival distribution for individuals with a certain combination of covariates.

One strategy would be to create a subset or stratum using individuals with a particular set of covariates, or at least in a range of values if we are considering continuous covariates, and then estimate the survival distribution for this particular stratum by using a Kaplan-Meier estimate. If the number of variables is large or we choose a narrow range, then the number of individuals in any subset would be so small as to make the Kaplan-Meier useless, *i.e.*, if a subset contained just a few censored survival times, then the corresponding Kaplan-Meier estimator would be very imprecise.

In a proportional hazards model, we assume a certain structure in the relationship of the covariates to the hazard rate. Assuming the model is an adequate representation of the true relationship, we can take advantage of this structure to obtain better estimate of the survival distribution as a function of the covariates.

The proportional hazards model assumes that the relationship of the hazard rate at time t given the covariate Z, where Z is a q dimensional vector of covariates $(Z_1, ..., Z_q)^T$, is given by

$$\lambda(t|Z) = \lambda_0(t)\exp(\beta^T Z).$$

If the model is correct, then the hazard at time t for an individual whose covariate vector is

$Z = z^* = (z_1^*, ..., z_q^*)^T$ is

$$\lambda(t|Z = z^*) = \lambda_0(t)\exp(\beta^T z^*).$$

Note: We use z^* here to emphasize that we are particularly interested in the survival function for a randomly sampled subject with this particular covariate. It should not be confused with the covariate values for the subjects in the study sample, where we will use Z_i to indicate the covariate values for subject i.

Because of the relationship of hazard to survival, this would imply that the survival distribution given $Z = z^*$ is

$$S(t|Z = z^*) = \mathrm{e}^{-\Lambda(t|Z=z^*)},$$

where $\Lambda(t|Z = z^*)$ is the cumulative hazard function given $Z = z^*$, *i.e.*,

$$\Lambda(t|Z = z^*) = \int_0^t \lambda(u|Z = z^*)du.$$

For the proportional hazards model,

$$\begin{aligned}
\Lambda(t|Z = z^*) &= \int_0^t \lambda(u|Z = z^*)du \\
&= \int_0^t \lambda_0(u)\exp(\beta^T z^*)du \\
&= \exp(\beta^T z^*)\int_0^t \lambda_0(u)du \\
&= \exp(\beta^T z^*)\Lambda_0(t),
\end{aligned}$$

where $\Lambda_0(t)$ is the cumulative baseline hazard function; *i.e.*,

$$\Lambda_0(t) = \int_0^t \lambda_0(u)du.$$

Consequently, the survival function for given $Z = z^*$ is

$$S(t|Z = z^*) = \mathrm{e}^{-\exp(\beta^T z^*)\Lambda_0(t)}.$$

This means that in order to estimate $S(t|Z = z^*)$, we only need to estimate β and $\Lambda_0(t)$. The parameter β can be estimated by MPLE $\hat{\beta}$ from the partial likelihood. So we only need to

get an estimate $\hat{\Lambda}_0(t)$ for $\Lambda_0(t)$. Then the estimate of $S(t|Z = z^*)$ would be given by

$$\hat{S}(t|Z = z^*) = \mathrm{e}^{-\exp(\hat{\beta}^T z^*)\hat{\Lambda}_0(t)}.$$

Note: We could choose any combination of the covariates z^* and find the corresponding estimate for the survival distribution for such a z^*.

Caution: Of course, all of this is predicated on the assumption that the proportional hazards model is an adequate representation to the data structure. We would not try to extrapolate these results to combinations of the covariates outside the range of the data even if the proportional hazards model was a reasonable fit the observed data.

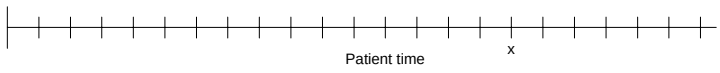
We are left with task of finding a reasonable estimate for the cumulative hazard function

$$\Lambda_0(t) = \int_0^t \lambda_0(u)du.$$

The logic for finding an estimate for the cumulative hazard function $\Lambda_0(t)$ in a proportional hazards model will be similar to that which was used in Chapter 2 to derive the Nelson-Aalen estimate of the cumulative hazard function in the one sample problem.

Recall that we divided the time axis into a grid of points using increasingly fine partition:

Figure 9.1: *Partition of time axis*



In the one-sample problem, all individuals in the sample have the same hazard of failing, implying the same cause-specific hazard (non-informative censoring). An estimate of $\lambda(x)\Delta x$ was obtained by using

$$\frac{dN(x)}{Y(x)} = \frac{\#\text{ of individuals in sample observed to die in } [x, x + \Delta x)}{\#\text{ of individuals in sample at risk at time } x}.$$

Since

$$\Lambda(t) \approx \sum_{x<t} \lambda(x)\Delta x.$$

This led us to the Nelson-Aalen estimate for $\Lambda(t)$:

$$\hat{\Lambda}(t) = \sum_{x<t} \frac{dN(x)}{Y(x)}.$$

In a proportional hazards model, the individuals in the sample do not have the same hazard of failing at time x but rather have a hazard which depends on their covariate values. That is, the ith individual with covariate values $Z_i = (z_{i1}, ..., z_{iq})^T$, has hazard

$$\lambda_i(t) = \lambda_0(t)\exp(\beta^T Z_i).$$

Consequently, if we define the past history of failures, censoring, and covariates before time x, by $\mathcal{F}(x)$, then

$$\begin{aligned} \mathrm{E}\left[dN_i(x)\,|\mathcal{F}(x)\right] &= Y_i\lambda_i(x)\Delta x \\ &= \lambda_0(x)\exp(\beta^T z_i)Y_i\Delta x. \end{aligned}$$

Now $dN(x) = \sum_{i=1}^n dN_i(x)$ is the number of deaths in $[x, x+\Delta x)$ for our sample, and

$$\begin{aligned} \mathrm{E}\left[dN(x)\,|\mathcal{F}(x)\right] &= \mathrm{E}\left[\sum_{i=1}^n dN_i(x)\,|\mathcal{F}(x)\right] \\ &= \sum_{i=1}^n \mathrm{E}\left[dN_i(x)\,|\mathcal{F}(x)\right] \\ &= \sum_{i=1}^n \lambda_i(x)Y_i\Delta x \\ &= \lambda_0(x)\Delta x\sum_{i=1}^n \exp(\beta^T Z_i)Y_i(x). \end{aligned}$$

Therefore it seems reasonable to estimate $\lambda_0(x)\Delta x$ by using

$$\frac{dN(x)}{\sum_{i=1}^n \exp(\beta^T Z_i)Y_i(x)}.$$

Hence if we wanted to estimate

$$\Lambda_0(t) \approx \sum_{x<t} \lambda_0(x)\Delta x,$$

we would use

$$\sum_{x<t} \left[\frac{dN(x)}{\sum_{i=1}^n \exp(\beta^T Z_i) Y_i(x)} \right].$$

Note: If all the β's were equal to zero (*i.e.*, no relationship of hazard to the covariates), then the previous formula would reduce to

$$\sum_{x<t} \left[\frac{dN(x)}{Y(x)} \right],$$

giving us back the Nelson-Aalen estimator.

The property that the estimate

$$\sum_{x<t} \left[\frac{dN(x)}{\sum_{i=1}^n \exp(\beta^T Z_i) Y_i(x)} \right].$$

is approximately unbiased for $\Lambda_0(t)$ follows form the following logic similar to that used for the Nelson-Aalen estimator.

$$\begin{aligned} & \mathrm{E}\left[\sum_{x<t} \left(\frac{dN(x)}{\sum_{i=1}^n \exp(\beta^T Z_i) Y_i(x)} \right) \right] \\ = & \sum_{x<t} \mathrm{E}\left[\frac{dN(x)}{\sum_{i=1}^n \exp(\beta^T Z_i) Y_i(x)} \right] \\ = & \sum_{x<t} \mathrm{E}\left[\mathrm{E}\left[\frac{dN(x)}{\sum_{i=1}^n \exp(\beta^T Z_i) Y_i(x)} \middle| \mathcal{F}(x) \right] \right]. \end{aligned}$$

In the inner expectation, the denominator

$$\sum_{i=1}^n \exp(\beta^T Z_i) Y_i(x)$$

is fixed conditional on $\mathcal{F}(x)$, therefore, the inner expectation is equal to

$$\begin{aligned} \mathrm{E}\left[\frac{dN(x)}{\sum_{i=1}^n \exp(\beta^T Z_i) Y_i(x)} \middle| \mathcal{F}(x) \right] &= \frac{\mathrm{E}[dN(x)|\mathcal{F}(x)]}{\sum_{i=1}^n \exp(\beta^T Z_i) Y_i(x)} \\ &= \frac{\mathrm{E}\left[\sum dN_i(x)|\mathcal{F}(x)\right]}{\sum_{i=1}^n \exp(\beta^T Z_i) Y_i(x)} \\ &= \frac{\lambda_0(x)\Delta x \sum \exp(\beta^T Z_i) Y_i(x)}{\sum_{i=1}^n \exp(\beta^T Z_i) Y_i(x)} \\ &= \lambda_0(x)\Delta x. \end{aligned}$$

Since $\lambda_0(x)\Delta x$ is not a random variable, the outer expectation is also $\lambda_0(x)\Delta x$. Consequently, the total expectation is

$$\mathrm{E}\left[\sum_{x<t}\left(\frac{dN(x)}{\sum_{i=1}^{n}\exp(\beta^T Z_i)Y_i(x)}\right)\right]=\sum_{x<t}\lambda_0(x)\Delta x\approx\Lambda_0(t).$$

The formula above involves the parameter vector β which also needs to be estimated from the data. Substituting the MPLE $\hat{\beta}$ yields an estimator for the cumulative baseline hazard function given by

$$\hat{\Lambda}_0(t)=\sum_{x<t}\left[\frac{dN(x)}{\sum_{i=1}^{n}\exp(\hat{\beta}^T Z_i)Y_i(x)}\right],$$

which is referred to as the Breslow estimator, Breslow (1972).

Therefore, if we wanted to estimate the survival function for an individual with covariate vector $z^*=(z_1^*,...,z_q^*)^T$, we could use

$$\hat{S}(t|z^*)=\exp\left[-\hat{\Lambda}_0(t)\exp(\hat{\beta}^T z^*)\right].$$

Standard error for $\hat{S}(t|z^*)$ and confidence intervals for $\hat{S}(t|z^*)$ could also be obtained. These formula are a bit complex and will be derived in the Advanced Survival Analysis class. The large sample properties for $\hat{S}(t|z^*)$ were derived by Tsiatis (1981) and by the use of counting processes by Andersen and Gill (1982).

Survival estimates for the proportional hazards model are given in `SAS`. Enclosed is an example we consider nodal status and ER status for `CALGB 8082` data. This is only for illustrative purposes. A more complete analysis should include all important prognostic factors

Appendix: SAS program

The following is a `SAS` program to estimate survival functions for some combinations of covariate:

```
options ps=72 ls=72;

data bcancer;
  infile "cal8082.dat";
  input days cens trt meno tsize nodes er;
  trt = trt - 1;
```

```
  label days="(censored) survival time in days"
    cens="censoring indicator"
    trt="treatment"
    meno="menopausal status"
    tsize="size of largest tumor in cm"
    nodes="number of positive nodes"
    er="estrogen receptor status"
    trt="treatment indicator";
run;

data bcancer1; set bcancer;
  if nodes = 1 or nodes = 10;
  if nodes=1 and er=0 then cat=1;
  if nodes=1 and er=1 then cat=2;
  if nodes=10 and er=0 then cat=3;
  if nodes=10 and er=1 then cat=4;
  if nodes=. then delete;
run;

data covars;
  input nodes er;
  cards;
  1 0
  1 1
  10 0
  10 1
;

title "Get survival estimate for each combination of nodes and er";
proc phreg data=bcancer;
  model days*cens(0) = nodes er;
  baseline out=a covariates=covars survival=s/nomean;
run;

data a1; set a;
  if nodes=1 and er=0 then cat=1;
  if nodes=1 and er=1 then cat=2;
  if nodes=10 and er=0 then cat=3;
  if nodes=10 and er=1 then cat=4;
run;

title "KM estimates for each category";
proc lifetest plots=(s) graphics notable data=bcancer1;
  time days*cens(0);
  strata nodes er;
  symbol1 v=none color=black line=1;
  symbol2 v=none color=black line=2;
  symbol3 v=none color=black line=3;
  symbol4 v=none color=black line=4;
run;

title "Survival estimates for each category";
proc gplot data=a1;
  plot s*days=cat;
  symbol1 interpol=join color=black line=1;
  symbol2 interpol=join color=black line=2;
  symbol3 interpol=join color=black line=3;
  symbol4 interpol=join color=black line=4;
run;

data _null_; set bcancer1;
  file "cat.dat";
```

```
  put days cens cat;
run;

data _null_; set a1;
  file "estsurv.dat";
  put days s cat;
run;
```

The following two graphes are generated using the following r functions:

```
postscript(file="estsurv1.ps", horizontal = F, height=6, width=8.5)
par(mfrow=c(1,2))

dat <- read.table(file="cat.dat", col.names=c("days", "cens", "cat"))
fit <- survfit(Surv(days, cens) ~ cat, dat)
plot(fit, xlab="Survival time in days", ylab="Survival probabilities",
lty=c(1,2,3,4))

dat <- read.table(file="estsurv.dat", col.names=c("days", "sprob", "cat"))
plot(0,0, xlab="Survival time in days", ylab="Survival probabilities",
pch=" ", xlim=c(0, max(dat$days)), ylim=c(0,1))

for (i in 1:4){
  lines(dat$days[dat$cat==i],dat$sprob[dat$cat==i], lty=i)}
legend(10, 0.2, c("cat 1", "cat 2", "cat 3", "cat 4"), lty=1:4, cex=0.8)
dev.off()
```

Figure 9.2: *KM estimate (left) and estimated survival curve (right) for each category*