When are PH, AFT and PO Models not Adequate for Health Risk Assessment?

Sujit K. Ghosh

NC State University
DEPARTMENT OF STATISTICS

http://www.stat.ncsu.edu/people/ghosh/
sujit.ghosh@ncsu.edu

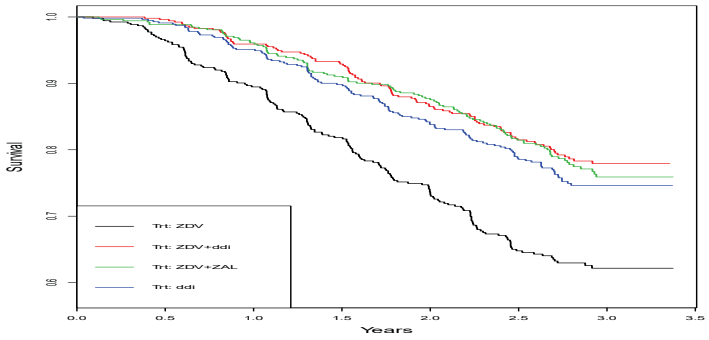
Presented at:

SAMSI GDRR Opening Workshop

SAS Hall, NC State University, Raleigh

https://tinyurl.com/samsi-gdrr

Outline

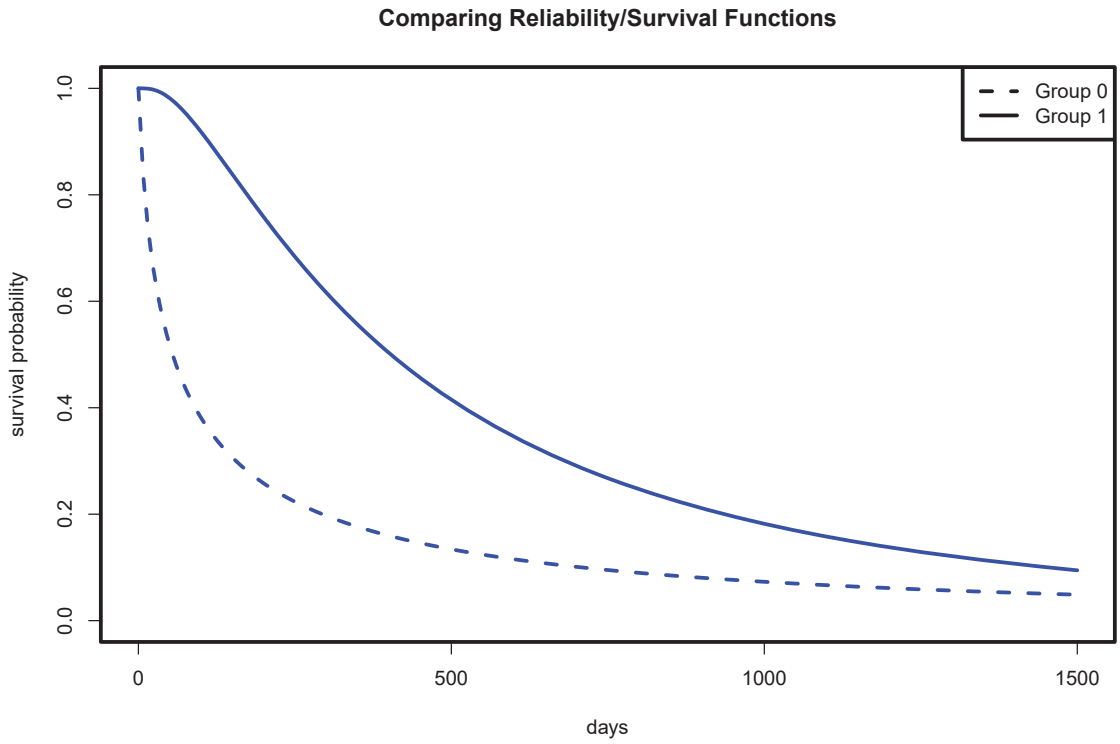


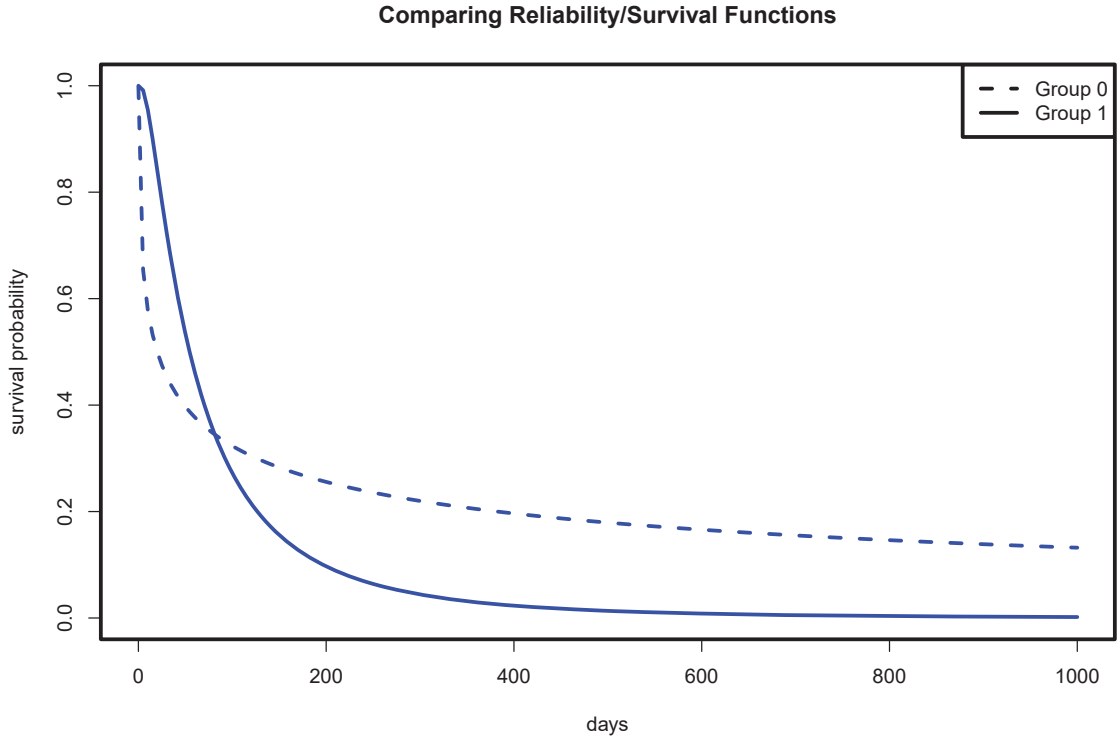
- Why/When are PHs, POs & AFT models not adequate?
- Conditional Models for Censored Data
- Theoretical Properties
- Numerical Illustrations
- Concluding Remarks and Open Questions

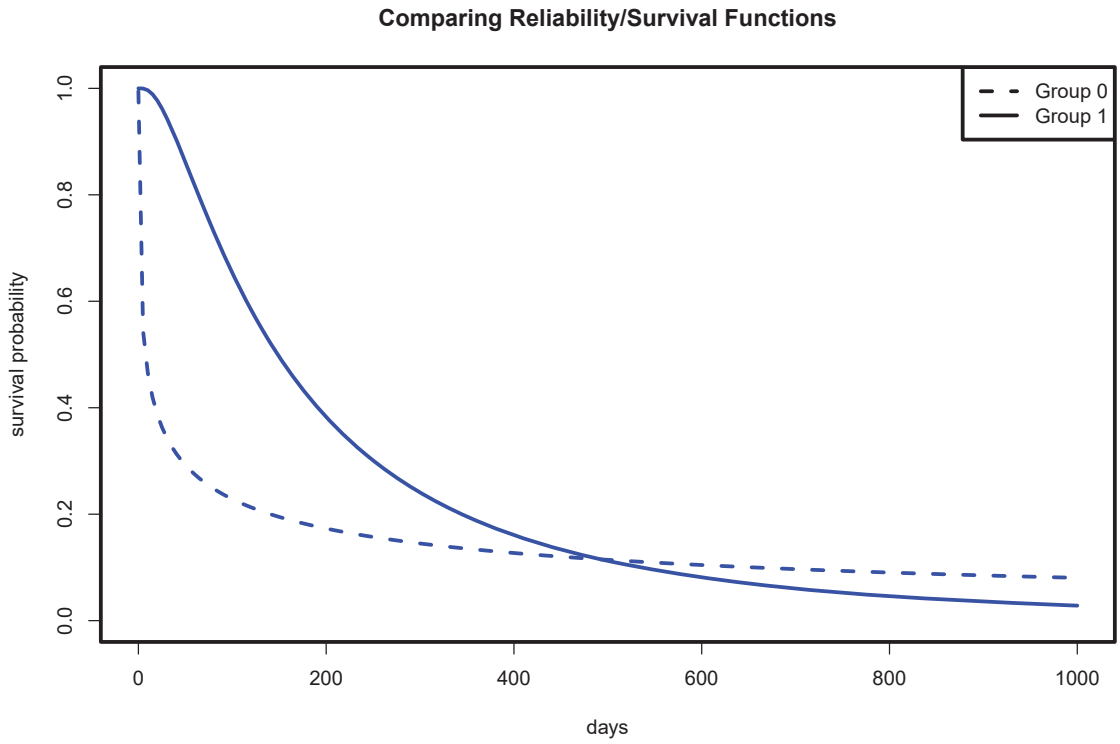
Why/When PHs, POs & AFTs are not adequate?

- Consider a gastric cancer study [Stablein et al. (1981)]
- A total of $n = 90$ patients with locally advanced gastric carcinoma were randomized to two treatment groups:
  - one group ($z = 0$) received only chemotherapy (45 patients); and
  - another group ($z = 1$) received radiotherapy together with the same chemotherapy
- The study was followed over eight years (and hence censored)
- How would you evaluate (health) risks between the two groups?
- It is common to compare the reliability or survival functions by evaluating the chance that a patient survives beyond a given unit of time

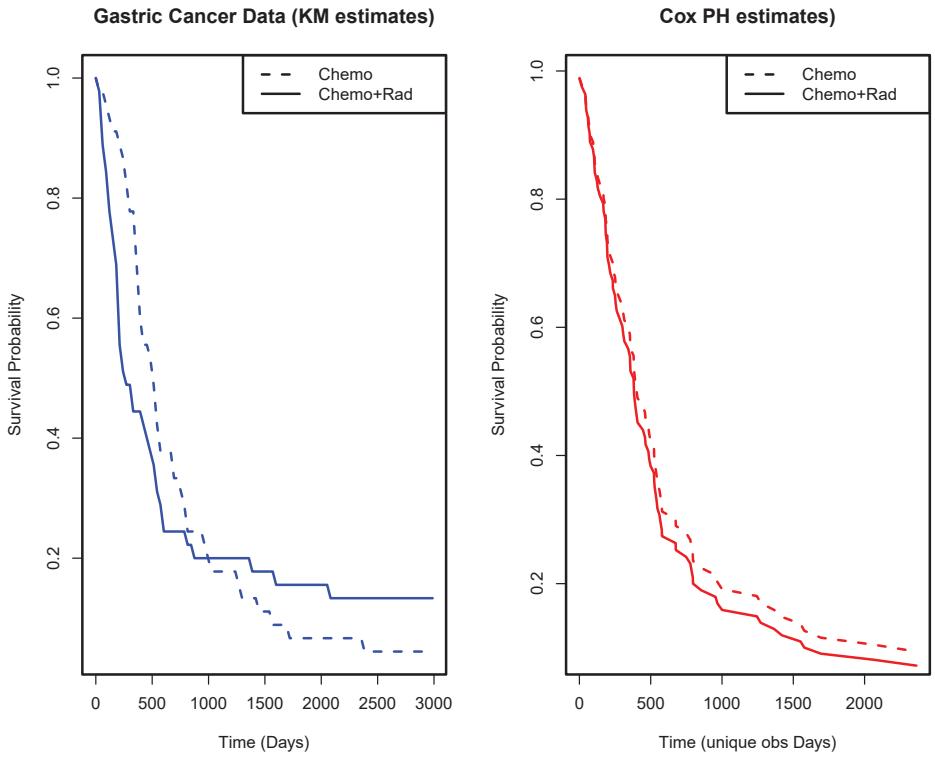
- Let $z = 0$ denote group receiving chemotherapy
  and $z = 1$ denote the group receiving radiotherapy together with chemotherapy
- Let $T =$ time to cancer remission since study entry
- Let $S_0(t) = \Pr[T > t|z = 0]$: survival function for chemo group
  and $S_1(t) = \Pr[T > t|z = 1]$: survival function for chemp+radio group
- The goal is to find if $S_0(t)$ is *'different'* from $S_1(t)$ using a suitable metric
  (e.g. $D^+(a, b) = \int_a^b (S_1(t) - S_0(t))^+ dt$ for some $[a, b] \subseteq (0, \infty)$)
- Suppose the study is censored at a time C and we observe only censored time $Y = \min(T, C)$ and its censoring indicator $\Delta = \mathbb{I}(T \leq C)$
- Thus, for each subject $i = 1, 2, \ldots, n$, we observe the triplet (Y_i, Δ_i, z_i) where $Y_i = \min(T_i, C_i)$ and $\Delta_i = \mathbb{I}(T_i \leq C_i)$
- Assume $(T_i, C_i, Z_i) \overset{iid}{\sim} (T, C, Z)$ for $i = 1, \ldots, n$ where $T \perp C|Z$


Comparing Reliability/Survival Functions
Group 0
Group 1
survival probability
0.0
0.2
0.4
0.6
0.8
1.0
0
500
1000
1500
days


Comparing Reliability/Survival Functions
Group 0
Group 1
survival probability
0.0
0.2
0.4
0.6
0.8
1.0
0
200
400
600
800
1000
days

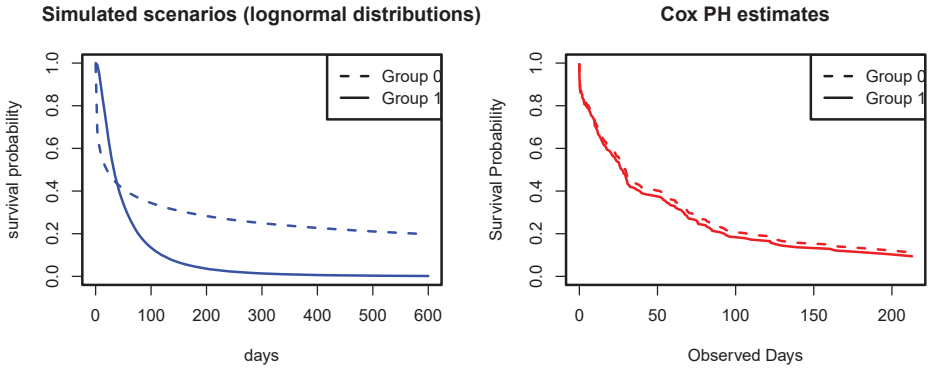


- How do we model the conditional survival function $S(t|z)$?
- Three popular models are based on conditional hazard function $h(t|z) = \frac{\partial}{\partial t}(-\log S(t|z))$. Let $h_j(t) = h(t|z = j)$ for $j = 0, 1$
  - (i) Proportional hazards (PH): $h_1(t) = \eta h_0(t)$ for some $\eta > 0$
    Equivalently, $S_1(t) = S_0(t)^\eta$
  - (ii) Accelerated Failure Time (AFT): $h_1(t) = \eta h_0(\eta t)$ for some $\eta > 0$
    Equivalently, $S_1(t) = S_0(t\eta)$
  - (iii) Proprtional Odds (PO): $\frac{1-S_1(t)}{S_1(t)} = \eta \frac{1-S_0(t)}{S_0(t)}$ for some $\eta > 0$
- When $S_0(t) \neq S_1(t)$, there are only two possibilities for these three models:
  - (a) Either $\eta > 1$, and then $S_1(t) < S_0(t)$ for all $t > 0$
  - (b) Or $\eta < 1$, and then $S_1(t) > S_0(t)$ for all $t > 0$
- Thus, **any of these three models will NOT allow the possibility of crossing survival functions**, i.e., $\nexists\, t_0 > 0$ such that $S_1(t_0) = S_0(t_0)$

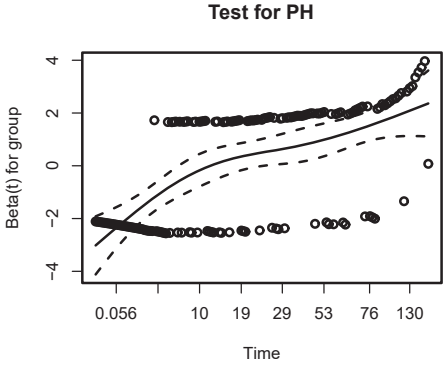


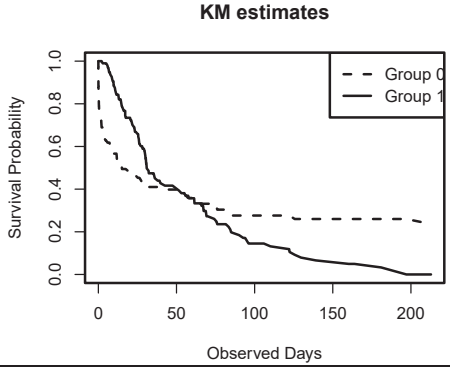
Consider the following simulated scenarios:

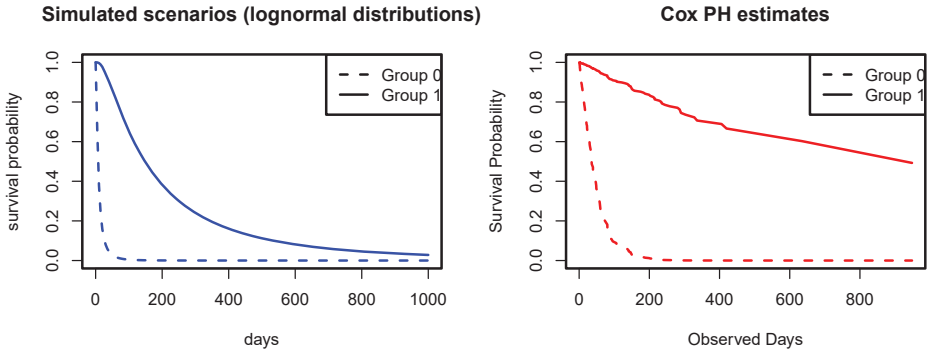
- Suppose $T|Z = z \sim LogNorm(\mu_z, \sigma_z)$ where $z \in \{0, 1\}$
- Suppose $C \sim Exp(\lambda)$ with mean $1/\lambda$
- We observe $Y = \min(T, C)$ for $z \in \{0, 1\}$ and $\Delta = \mathbb{I}(T \leq C)$
- Observed data: $\{(Y_i, \Delta_i, Z_i)\}$ for $i = 1, \ldots, n$
- Obtain estimates of the survival/reliability functions: $S(t|z)$ for $z = 0, 1$
  (i) Using Kaplan-Meier estimates separately for each group
  (ii) Assuming proportional hazard: $S(t|z = 1) = S(t|z = 0)^{\eta}$ for some $\eta > 0$
- Consider two scenarios with $\lambda = 1/400$ and $n = 100$:
  - Case 1: $\mu_0 = 3, \sigma_0 = 4, \mu_1 = 3.5, \sigma_1 = 1$
  - Case 2: $\mu_0 = 2, \sigma_0 = 1, \mu_1 = 5, \sigma_1 = 1$ (AFT with $\eta = e^{-3}$)


Simulated scenarios (lognormal distributions)
survival probability
Group 0
Group 1
days

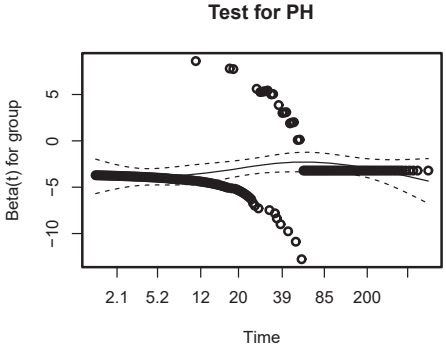
Cox PH estimates
Survival Probability
Group 0
Group 1
Observed Days

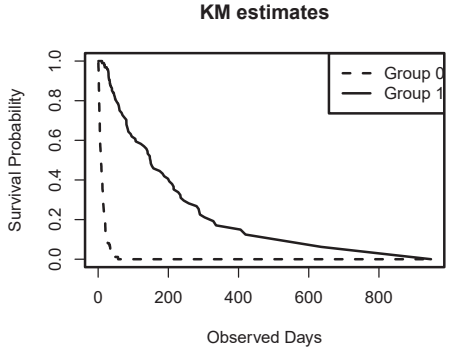

Test for PH
Beta(t) for group
Time


KM estimates
Survival Probability
Group 0
Group 1
Observed Days


Simulated scenarios (lognormal distributions)
survival probability
Group 0
Group 1
days

Cox PH estimates
Survival Probability
Group 0
Group 1
Observed Days


Test for PH
Beta(t) for group
Time


KM estimates
Survival Probability
Group 0
Group 1
Observed Days

Consider a more general set-up with vector valued covariate Z

- T: survival time for a patient with baseline covariate (vector) Z
- C: censoring time (typically independent of T) given Z
- $Y = \min(T, C)$: Observed survival time
- $\Delta = I(T \leq C)$: Observed censoring indicator
- Observe data: $\{(Y_i, \Delta_i, Z_i),\ i = 1, \ldots, n\}$
- Our goal is to estimate the conditional survival function: $S(t|z) = \Pr[T > t|Z = z]$ based on the observed data
- A reasonable assumption is that $(T_i, C_i, Z_i) \overset{iid}{\sim} (T, C, Z)$ where we further assume that $T \perp C \mid Z$
- Often further simplifying assumptions are made to estimate $S(t|z)$ or equivalently conditional hazard function $h(t|z) = \frac{\partial}{\partial t}(-\log S(t|z))$

- Three of the most popular models:

  (i) Proportional Hazard (PH) model: $h(t|z) = h_0(t)\eta(z^\intercal\beta)$ or equivalently $S(t|z) = S_0(t)^{\eta(z^\intercal\beta)}$ for some baseline survival function $S_0(t)$

  (ii) Accelerated Failure Time (AFT) model: $S(t|z) = S_0(t\eta(z^\intercal\beta))$

  (iii) Proportional Odds (PO) model: $\frac{1-S(t|z)}{S(t|z)} = \eta(z^\intercal\beta)\frac{1-S_0(t)}{S_0(t)}$

  where $\eta(\cdot)$ is non-negative increasing function with $\eta(0) = 1$ (e.g., $\eta(u) = e^u$)

- Notice that in the simplest case with $z \in \{0, 1\}$:

  (i) PH: $S_1(t) = S_0(t)^\eta$ (ii) AFT: $S_1(t) = S_0(t\eta)$ (iii) PO: $\frac{1-S_1(t)}{S_1(t)} = \eta\frac{1-S_0(t)}{S_0(t)}$

  where $S_1(t) = S(t|z = 1)$ and $S_0(t) = S(t|z = 0)$

  (a) None of these models allows for crossing survival functions

  (b) Even when survival functions don't cross but if we fit a PH model to an AFT model we get biased estimates of survival functions

- Hence, there is a need to develop more flexible models for $S(t|z)$

- Many extensions are available but often such models are computationally not as efficient as the PH model
- The most popular extension is to include a time-varying effect $\beta(t)$ by replacing β within the PH model
- Most of the methodologies involving the time-varying effect may turn out to be computationally intensive
- Clearly, *the appealing feature of easy interpretation and estimation of the PH model comes from the separation of time and covariate effects*
- Once such separation (and hence interpretation and simpler estimation) is lost, why should we insist on (time-varying) PH structure at all?
- Another important extension in this line of research is referred to as HARE (Hazard Regression), where log of conditional hazard function is modeled as linear splines

Conditional Models for Censored Data

- We consider nonparametric hazard regression based on a sequence of basis functions for right-censored data: $\{(Y_i, \Delta_i, Z_i); i = 1, \ldots, n\}$
- First, we consider the one-sample right-censored data with no covariates
- Following standard practice, assume that $\tau = \inf\{t : S(t) = 0\} < \infty$
- Notice that likelihood contribution of i-th observation (Y_i, Δ_i) is $\Delta_i \log h(Y_i) - H(Y_i)$ where $H(t) = \int_0^t h(s)$ is the cumulative hazard function
- Thus, it is sufficient to model the hazard function $h(t)$
- For $m = 2, 3, \ldots$, we approximate $h(\cdot)$ by a sieve of basis functions:

$$h_m(t, \boldsymbol{\gamma}) = \sum_{k=1}^{m} \gamma_k g_{m,k}(t) = \boldsymbol{\gamma}^\intercal \boldsymbol{g}_m(t), \; 0 \leq t < \infty, \tag{1}$$

- The pre-specified known basis functions $\boldsymbol{g}_m(t) = (g_{m,1}(t), ..., g_{m,m}(t))^\intercal$ satisfy $g_{m,k}(\cdot) \geq 0$ for $k = 1, \ldots, m$
- The coefficients $\boldsymbol{\gamma}_m = (\gamma_1, \gamma_2, ..., \gamma_m)^\intercal$ satisfy $\gamma_k \geq 0, \ \forall k, m$
- The corresponding cumulative hazard function is given by

$$H_m(t, \boldsymbol{\gamma}) = \sum_{k=1}^{m} \gamma_k G_{m,k}(t) = \boldsymbol{\gamma}^\intercal \boldsymbol{G}_m(t), \ 0 \leq t < \infty, \tag{2}$$

  where $\boldsymbol{G}_m(t) = (G_{m,1}(t), ..., G_{m,m}(t))^\intercal$ with $G_{m,k}(t) = \int_0^t g_{m,k}(u)du$
- Clearly, the monotonicity of $H_m(\cdot)$ is enforced by the restriction that $\gamma_k \geq 0$ and $g_{m,k}(\cdot) \geq 0$ for $k = 1, 2, ..., m$
- *As both hazard and cumulative hazard functions are linear in unknown coefficient $\boldsymbol{\gamma}$, the model will shown to have theoretical and computational advantages*
- Next, we choose a sequence of basis functions that provides uniform convergence

- We use the sequence of Bernstein basis functions:
  $g_{m,k}(t) = \frac{m}{\tau}\binom{m-1}{k-1}(\frac{t}{\tau})^{k-1}(1 - \frac{t}{\tau})^{m-k}\mathbb{I}(0 \leq t \leq \tau)$ for $m = 2, 3, \ldots$
  (which is the density of a scaled $Beta(k, m - k + 1)$ distribution)
- More specifically, for any continuous hazard function $h(t)$, we can show that

$$\max_{t \in [0,\tau]} |h_m(t, \boldsymbol{\gamma}) - h(t)| \to 0 \ \text{ as } m \to \infty$$

  if we choose $\gamma_k = h(\frac{k-1}{m-1}\tau)$ for $k = 1, \ldots, m$
- Notice that a legitimate hazard function besides being non-negative should also satisfy $\int_0^\infty h(t)dt = \infty$ (if we require $S(\infty) = 0$)
- Thus, to complete the model specification, we define

$$h_m(t, \boldsymbol{\gamma}) = \sum_{k=1}^{m} \gamma_k g_{m,k}(t)\mathbb{I}(0 \leq t \leq \tau) \ + \ \frac{m\gamma_m}{\tau}\mathbb{I}(t \geq \tau)$$

- Thus, it follows that the log-likelihood function of $\boldsymbol{\gamma}$ can be written as

$$\begin{aligned} l(\boldsymbol{\gamma}) &= \sum_{i=1}^{n} \{\Delta_i \log(h_m(Y_i, \boldsymbol{\gamma})) - H_m(Y_i, \boldsymbol{\gamma})\} \\ &= \sum_{i=1}^{n} \{\Delta_i \log(U_i^\intercal \boldsymbol{\gamma}) - V_i^\intercal \boldsymbol{\gamma}\}, \end{aligned} \tag{3}$$

  where $\boldsymbol{\gamma} \in \mathcal{C}_m = [0, \infty)^m$, $U_i = \boldsymbol{g}_m(Y_i)$, and $V_i = \boldsymbol{G}_m(Y_i)$

- Notice that the existence and uniqueness of the (sieve) maximum likelihood estimator follows from strict concavity of the above log-likelihood function

- Moreover, as the gradient and Hessian of the log-likelihood is available in closed forms, a modified quasi-Newton method can be easily implemented (`optim` in R)

- Next, we show that *the form of log-likelihood with covariates remains essentially the same as that with no covariates*

- For simplicity, consider again the two-groups case with $Z \in \{0, 1\}$

- The conditional hazard and cumulative hazard functions are given by

$$\begin{aligned} h_m(t, \boldsymbol{\gamma}|Z) &= \{(1-Z)\boldsymbol{\gamma_0} + Z\boldsymbol{\gamma_1}\}^\intercal \boldsymbol{g}_m(t) \text{ and} \\ H_m(t, \boldsymbol{\gamma}|Z) &= \{(1-Z)\boldsymbol{\gamma_0} + Z\boldsymbol{\gamma_1}\}^\intercal \boldsymbol{G}_m(t), \end{aligned} \tag{4}$$

  where $\boldsymbol{\gamma}^\intercal = (\boldsymbol{\gamma_0}^\intercal, \boldsymbol{\gamma_1}^\intercal) = (\gamma_{01}, \gamma_{02}, ..., \gamma_{0m}, \gamma_{11}, \gamma_{12}, ..., \gamma_{1m})^\intercal$

- Accordingly, the log-likelihood function again takes the form

$$l(\boldsymbol{\gamma}) = \sum_{i=1}^{n} \{\Delta_i \log(U_i^\intercal \boldsymbol{\gamma}) - V_i^\intercal \boldsymbol{\gamma}\}, \quad \text{where} \tag{5}$$

$$U_i = \begin{bmatrix} (1-Z_i)\boldsymbol{g}_m(X_i) \\ Z_i \boldsymbol{g}_m(X_i) \end{bmatrix} \text{ and } V_i = \begin{bmatrix} (1-Z_i)\boldsymbol{G}_m(X_i) \\ Z_i \boldsymbol{G}_m(X_i) \end{bmatrix}.$$

- Thus, the log-likelihood function in (5) is of the same form as in the case of one-sample data with no covariates (see eq. (3))

- The model described in (4) can be regarded as modeling the discretized hazard function using 1-way ANOVA
- As a result, it can be further extended to the cases when there are multiple categorical covariates and each may have more than 2 levels
- Similarly, it can be shown that the same form of the likelihood form is retained even when the covariates are continuous (see Osman & Ghosh, 2012)
- Thus, the above model formulation leads to a very convenient log-likelihood form which enjoys theoretical and computational advantages
- Hence the existence and uniqueness of maximum likelihood estimate $\hat{\gamma}$ follow by the strict concavity of the log-likelihood function
- Moreover, Fisher Information can be obtained from the Hessian given by:

$$\nabla^2 l(\boldsymbol{\gamma}) = -\sum_{i=1}^{n} \Delta_i \frac{U_i U_i^{\mathsf{T}}}{(U_i^{\mathsf{T}}\boldsymbol{\gamma})^2}$$

Theoretical Properties

- The consistency and the rate of the convergence are obtained using the Hellinger distance as the metric of choice

$$d(h_1, h_2) = \left\{ \int (\sqrt{p_{h_1}} - \sqrt{p_{h_2}})^2 d\mu \right\}^{1/2}, \tag{6}$$

  where $h_j = h_j(\cdot)$ denote hazard functions and p_{h_j}, the induced density ($j = 1, 2$)

- Asymptotic properties of $\hat{h}_m(\cdot)$ are established using the following boundness and smoothness of the true hazard function h_0:

  (I) $\tau = \inf\{t > 0 : \int_0^t h_0(u)du = \infty\} < \infty$.

  (II) $h_0(\cdot)$ *is continuous on* $[0, \tau]$ *and* $h_0(t) \geq \varepsilon$ *for all* $t \in [0, \tau]$ *for some* $\varepsilon > 0$

  (III) The first derivative denoted by $h_0^{(1)}(\cdot)$*, is Holder continuous with the exponent* α_0

- Hence the parameter space is given by

$$\Theta = \{h(\cdot) \in C[0,\tau] : h(\cdot) \text{ satisfies (I)-(III)}\}. \quad (7)$$

- The parameter space Θ is approximated by smaller finite dimensional space so called the sieve given by

$$\Theta_m = \left\{ h_m(t) = \sum_{k=1}^{m} \gamma_k g_{m,k}(t) : \boldsymbol{\gamma} = (\gamma_1, \gamma_2, ..., \gamma_m)^T \in [0, L_\gamma]^m \right\}, \quad (8)$$

Theorem 1. *(Consistency) Suppose the conditions (I)-(II) hold and the sieve Θ_m is defined as in (8), then $d(\hat{h}_{m,n}, h_0) \stackrel{a.s.}{\to} 0$ as $m, n \to \infty$.*

Theorem 2. *(Rate of Convergence) Suppose the conditions (I)-(III) hold and the sieve Θ_m is defined as in (8), if $m = o(n^\kappa)$ with $\kappa = \frac{2}{3+2\alpha_0}$, then $d(\hat{h}_{m,n}, h_0) = O_p(n^{-\frac{1+\alpha_0}{3+2\alpha_0}})$.*

The proofs of these two theorems are given in Osman and Ghosh (2012)

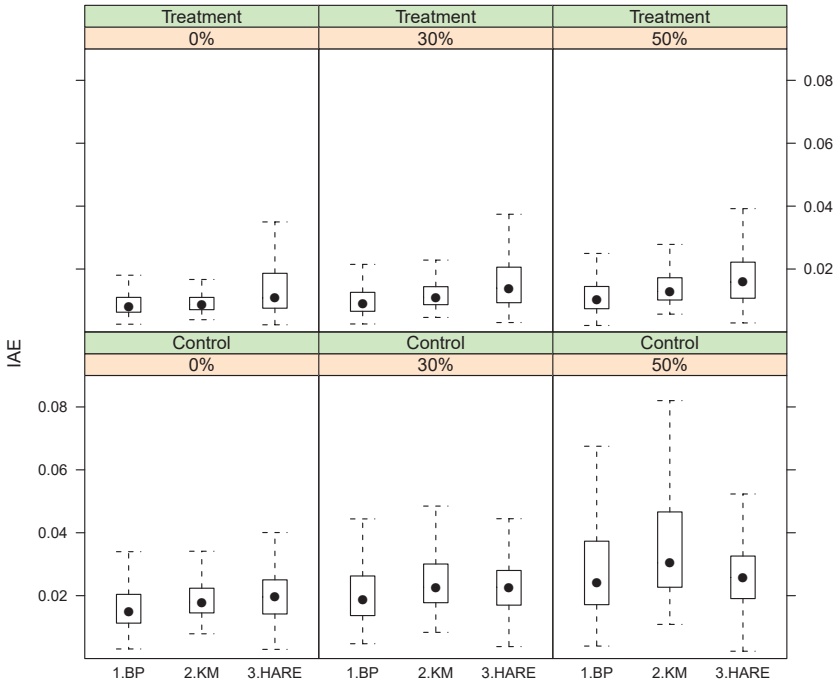
Numerical Illustrations

- We conducted several simulation studies to investigate the empirical performance of the proposed Bernstein polynomial based conditional hazard estimates
- Compared it with some popular nonparametric and semiparametric models
- We focused on the estimation of the survival function, which should give clinical practitioners more information under nonproportional hazards
- We used the integrated absolute error (IAE) defined by

$$IAE = \int_0^\tau |\hat{S}(t) - S_{true}(t)| dt$$

  where τ is the smallest value satisfying $S_{true}(\tau) \leq 0.001$
- Several practical scenarios were explored with varying rates of censoring
- Both categorical and continuous covariate cases were investigated

Two Group model: $T_0 \sim LN(-0.1, 0.5^2), T_1 \sim LN(0, 0.25^2), C \sim Exp(\lambda)$; $n_0 = n_1 = 50$; 1000 MC runs



	Control			Treatment		
	0%	30%	50%	0%	30%	50%
Median IAE ($\times 100$)						
BP	1.49	1.89	2.43	0.80	0.91	1.02
KM	1.79	2.26	3.05	0.87	1.10	1.29
HARE	1.96	2.27	2.58	1.08	1.39	1.59
Smallest IAE Achieved						
BP	64%	61%	54%	49%	64%	72%
KM	3%	2%	2%	25%	12%	6%
HARE	33%	37%	46%	26%	24%	22%

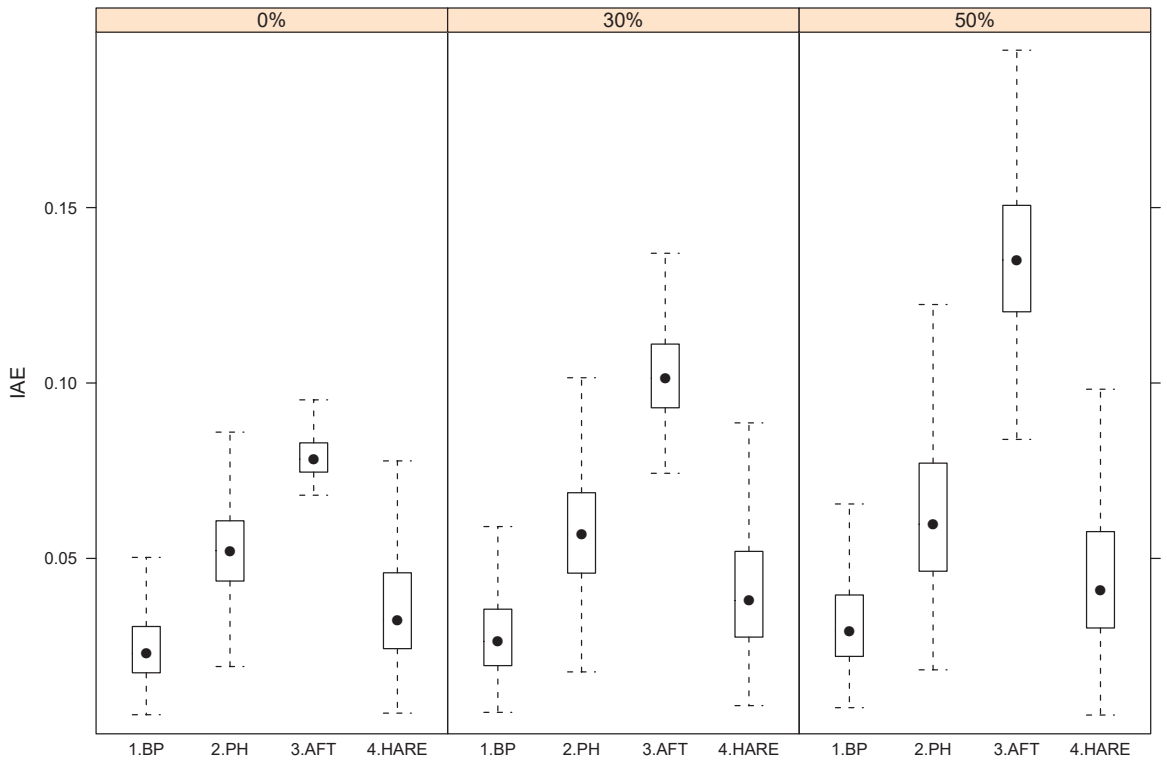
Continuous covariate case: The (T, Z) were generated by the following model:

$$\log T = \mu(Z) + \varepsilon,$$

where $\mu(Z) = \cos(\pi Z)$ and $\varepsilon | Z \sim N(0, \sigma(Z))$ with $\sigma(Z) = |Z|$ and $Z \sim U(0, 1)$

	Median IAE at $z = 0.5$ ($\times 100$)			Smallest IAE Achieved		
	0%	30%	50%	0%	30%	50%
BP	2.29	2.63	2.93	75%	76%	71%
PH	5.22	5.70	5.97	3%	6%	7%
AFT	7.83	10.13	13.51	0%	0%	0%
HARE	3.23	3.80	4.09	22%	19%	22%

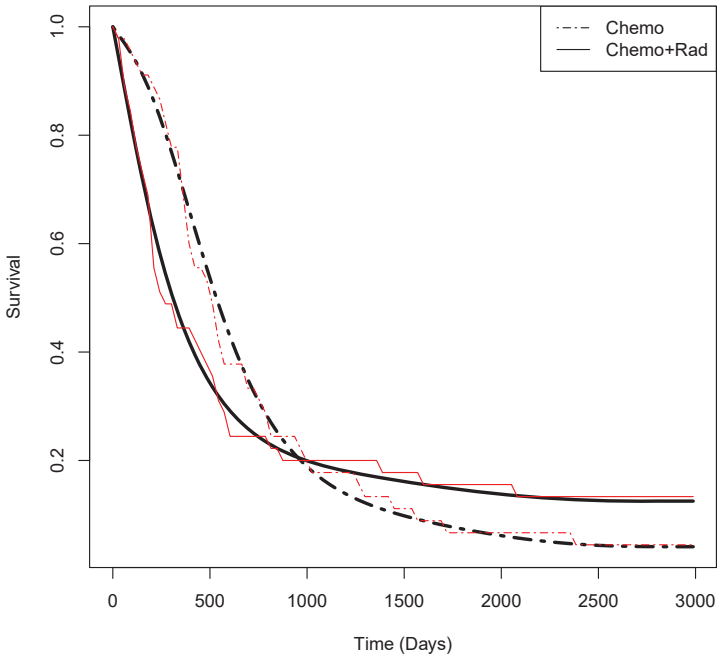
Remark: The PH and AFT models were fitted with the mean function $\mu(Z)$ misspecified as a linear function of Z. Also, the baseline distribution of the parametric AFT model is misspecified to be exponential distribution.



Analysis of Gastric Cancer Data

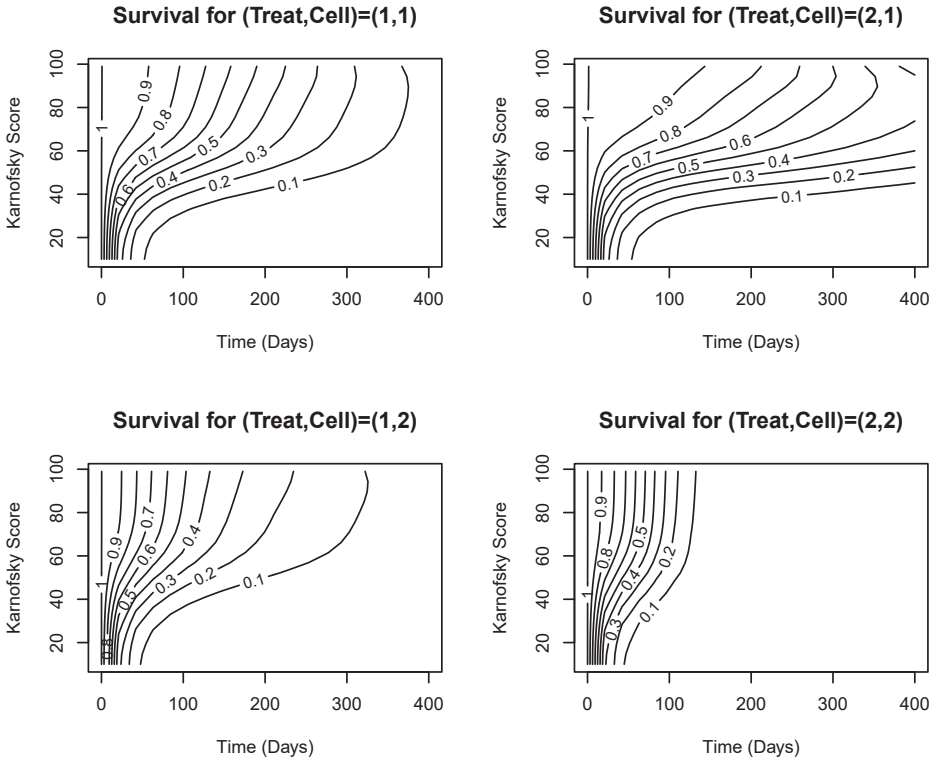
- Recall that $n = 90$ patients with locally advanced gastric carcinoma were randomized to two treatment groups (45 patients per group)
- One group only received chemotherapy while the other group received radiotherapy together with the same chemotherapy.
- As shown by the Kaplan-Meier curves, before the crossing point at approximately 1000 days the patients in the chemo group had better survival rates while the benefit of combination treatment of chemotherapy and radiotherapy started to emerge at a later stage of the study
- We estimated the survival functions using the BP model with the order $m = [n^{0.5}] = 10$ (i.e., assuming $\alpha_0 = 0.5$ in Theorem 2)
- The results indicate that the estimated smooth curves cross at $t = 952$ days

Estimated smooth survival curves along with KMEs:**Gastric Cancer Data**



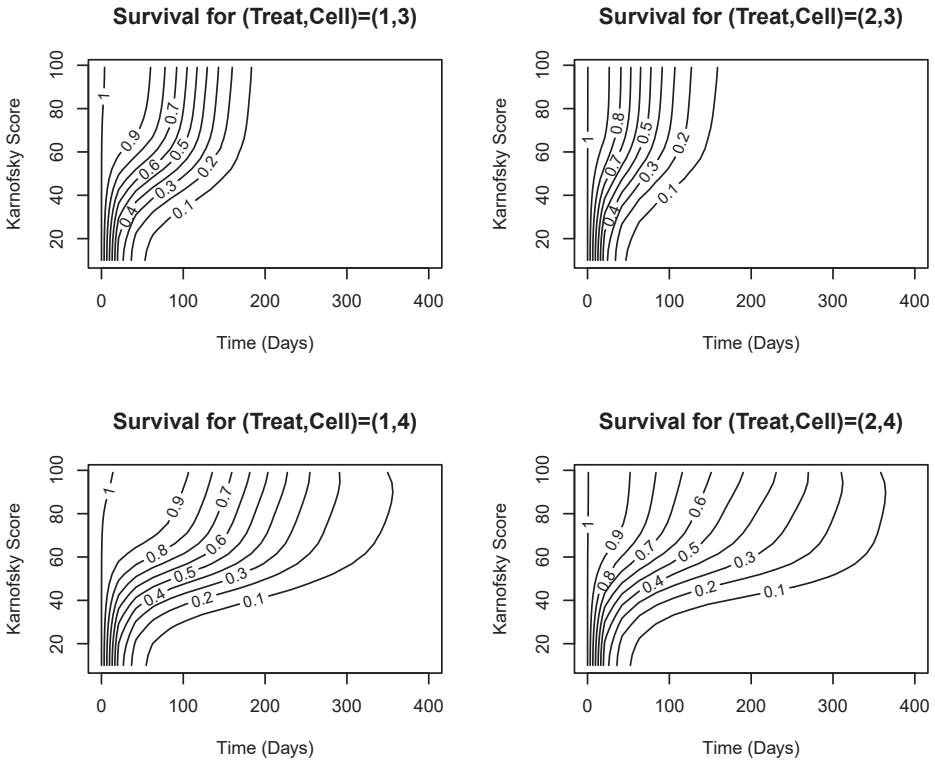
Analysis of Veterans Administration Data

- In the Veterans Administration lung cancer study $n = 137$ male patients were assigned to two treatment groups: standard chemotherapy and test chemotherapy
- In addition, 5 other baseline covariates were recorded
- Following previous works we only used Karnofsky performance score and cell type as important covariates
- Karnofsky score is a continuous variable that takes value from 0 to 100
- Cell type has fours levels: squamous cell, small cell, adenocarcinoma and large cell
- We analyzed the data using the partially linear coefficient model (see Osman and Ghosh, 2012) with the order $m = [n^{0.5}] = 12$



Survival for (Treat,Cell)=(1,3) Survival for (Treat,Cell)=(2,3)

(contd.)



- Generally, patients with higher Karnofsky scores have higher survival rates
- But such association differs across treatment groups and cell types
- Overall, the patients with small cell type in the test treatment group underwent the sharpest decline in survival rates
- While the patients with squamous cell type receiving the test chemotherapy had the best survival profiles among all groups
- Among the patients with small cell type, the patients receiving the standard chemotherapy appear to have better survival rates than those receiving the test chemotherapy
- On the contrary, the patients with squamous cell type, those receiving the test treatment had better survival rates than the ones receiving the standard treatment
- For the patients with adenocarcinoma or large cell type, survival contours are similar across the treatment groups.

Concluding Remarks and Open Questions

- The most remarkable feature of the proposed method is that the log-likelihood, its gradient, and the Hessian matrix all take a relatively simple form
- Additionally, we show that the general simple form of the log-likelihood function holds even in the presence of categorical and continuous covariates
- Under some mild conditions, the proposed sieve maximum likelihood estimator is shown to be consistent and the corresponding rate of convergence is obtained
- The proposed method provides similar or slightly better estimates than the HARE model but the proposed method has computational stability compared to HARE
- However...there are several open questions that remains to be answered...

- *How to perform a data driven choice of m that also maintains theoretical properties?*
- Recall that the Theorem 2 (slide# 20) specified $m = o(n^{\kappa})$ where $\kappa \in (0, 1)$ depends on smoothness of the underlying hazard function
- *Can we develop an adaptive method that wouldn't require knowing such smoothness order?*
- *How to extend both HARE and the proposed model to high-dimensional covariates that performs automatic variable selection?*
- It is worth noting that no regularization is needed when parameters are constrained to be non-negative; e.g., see the recent paper

Koike, Y. and Tanoue Y. (2019). Oracle inequalities for sign constrained generalized linear models, *Econometrics and Statistics*, **11**, 145-157: https://doi.org/10.1016/j.ecosta.2019.02.001

Questions?

Osman, M. and Ghosh, S. K. (2012). Nonparametric Regression Models for Right-censored Data using Bernstein Polynomials, *CSDA*, **56**, 559-573:
https://doi.org/10.1016/j.csda.2011.08.019

Kooperberg, C., Stone, C.J. and Truong, Y.K. (1995). Hazard Regression, *JASA*, **90**, 78-94:
https://www.jstor.org/stable/2291132

R packages: https://cran.r-project.org/web/views/Survival.html

R codes with examples

```
#Case 1:
mu0=3; sigma0=4; mu1=3.5; sigma1=1
#Case 2:
#mu0=2; sigma0=1; mu1=5; sigma1=1
S0=function(t){pnorm(log(t),mean=mu0,sd=sigma0,lower.tail=F)}
S1=function(t){pnorm(log(t),mean=mu1,sd=sigma1,lower.tail=F)}
#Plot true survival functions:
par(mfrow=c(2,2),lwd=2)

title1="Simulated scenarios (lognormal distributions)"
days=seq(0.001,1000,l=200)
plot(days, S1(days), type="l", col="blue", ylab="survival probability",
ylim=c(0, 1), main=title1)
lines(days, S0(days), lty=2, col="blue")
legend("topright", legend = c("Group 0", "Group 1"), lty = c(2,1))
```

```
#Generate samples and fit KM & PH curves:
n=100 #number of observation in each group
T0=exp(rnorm(n,mean=mu0,sd=sigma0)); T1=exp(rnorm(n,mean=mu1,sd=sigma1))
T=c(T0,T1); Cen=rexp(2*n,rate=1/400)
Y=pmin(T,Cen); Delta=as.numeric(T<=Cen); group=c(rep(0,n),rep(1,n))

library(survival)
fit.ph=coxph(Surv(Y,Delta)~group)
summary(fit.ph); eta.hat=exp(as.numeric(fit.ph$coef))
details.ph=coxph.detail(fit.ph)
obs.days=details.ph$time
S0.PH<-exp(-cumsum(details.ph$hazard)); S1.PH<-S0.PH^eta.hat

title2="Cox PH estimates"
plot(obs.days,S0.PH,type="l",lty=2,main=title2,ylab="Survival Probability",
 xlab="Observed Days", ylim=c(0, 1), col='red')
lines(obs.days,S1.PH,lty=1,col='red')
legend("topright", legend = c("Group 0", "Group 1"), lty = c(2, 1))
```



```
#Perform test for PH assumption
test.ph<-cox.zph(fit.ph)
print(test.ph); title3="Test for PH"; plot(test.ph, main=title3)

#Obtain KM estmates
library(splines)
fit.km=summary(survfit(Surv(Y,Delta)~group),times=obs.days,extend=TRUE)
m=length(obs.days)
S0.KM=fit.km$surv[1:m]; S1.KM=fit.km$surv[(m+1):(2*m)]

title4="KM estimates"
plot(obs.days,S0.KM,type="l",lty=2,main=title4,
ylab="Survival Probability",xlab="Observed Days", ylim=c(0,1))
lines(obs.days,S1.KM,lty=1)
legend("topright", legend = c("Group 0", "Group 1"), lty = c(2, 1))
```

```
#########################################################
#Unconditional hazard function estimation using BPs:#
#########################################################
BPsurv=function(y,d,m=10){
n=length(y); m=min(m,floor(n/log(n))); tau=max(y)
#hazard function:
h=function(t,gama){m=length(gama)
return(sum(gama*dbeta(t/tau,1:m,m:1)/tau)+as.numeric(t>tau)*m*gama[m]/tau)}
#Cumulative hazard function:
H=function(t,gama){m=length(gama)
return(sum(gama*pbeta(t/tau,1:m,m:1))+max(t-tau,0)*m*gama[m]/tau)}
#Negative log-likelihood function:
nloglik=function(gamma){
-sum(d*log(sapply(y,h,gama=gamma))-sapply(y,H,gama=gamma))}
fit=optim(par=rep(1,m),fn=nloglik, method="L-BFGS-B",lower=rep(1.0e-8,m))
gamma.est=fit$par; h.est=function(t){h(t,gama=gamma.est)}
hFun=function(t){sapply(t,h.est)}; H.est=function(t){H(t,gama=gamma.est)}
HFun=function(t){sapply(t,H.est)}; SFun=function(t){exp(-HFun(t))}
return(list(SFun=SFun,hFun=hFun,HFun=HFun,m=m))}
```