Inference about Conditional Hazard Function without PH and AFT Assumptions

Sujit K. Ghosh

NC State University
DEPARTMENT OF STATISTICS

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

Presented at:

2018 ASA Biopharmaceutical Section Statistics Workshop

Washington, DC, USA

http://ww2.amstat.org/meetings/biopharmworkshop/2018/

Outline

Two Equivalent Modeling Options:

$$\text{Conditional Survival function: } S(t|z) = \Pr[T > t|Z = z]$$

$$\text{Conditional Hazard function: } h(t|z) = \frac{\partial}{\partial t}(-\log(S(t|z)))$$

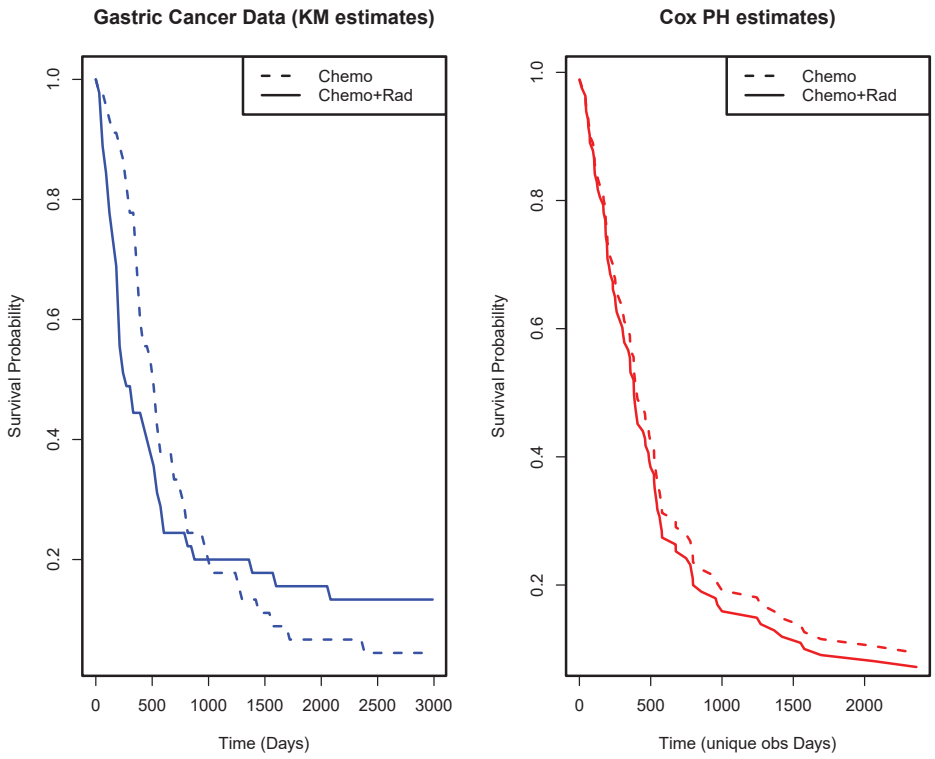
- Why/When are PHs, POs & AFT models not adequate?
- Conditional Models for Censored Data
- Two Real Case Studies
- Concluding Remarks
- Theoretical Properties (optional)

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
- Is there a difference between the two treatments?
- *Can we make inference using one of the popular models: proportional hazards (PHs), proportional odds (POs) or accelerated failure time (AFT) models?*

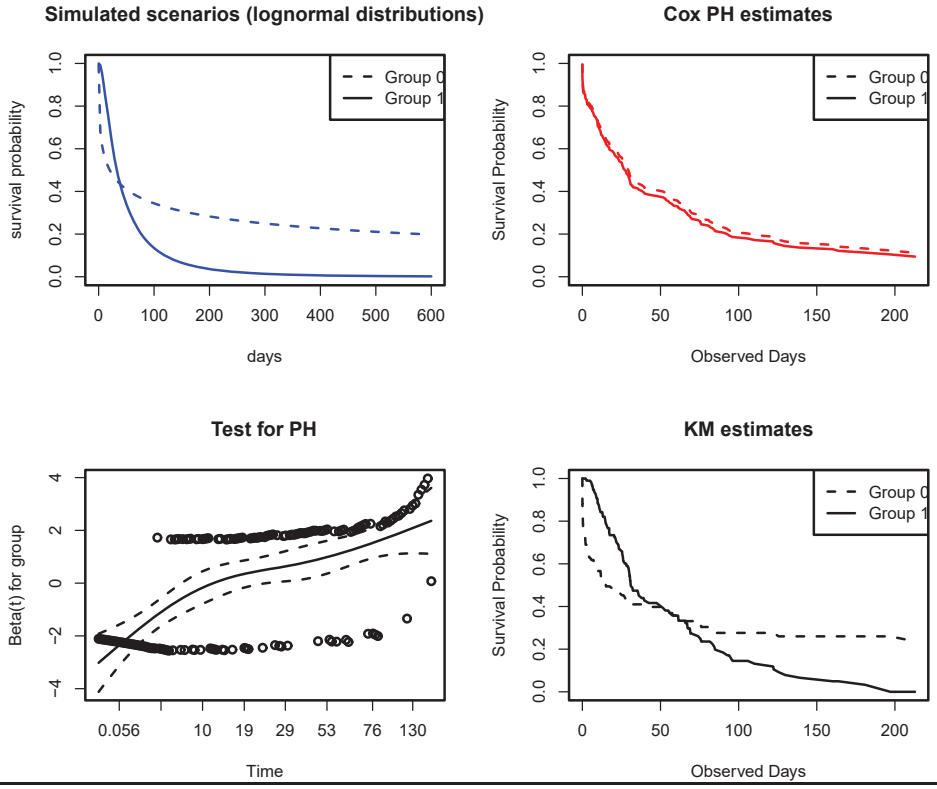
- Let $z = 0$ denote chemotherapy group and $z = 1$ denote the radiotherapy together with chemotherapy group
- Let $T =$ time to cancer remission since study entry
- Let $S_0(t) = \Pr[T > t|z = 0]$ and $S_1(t) = \Pr[T > t|z = 1]$ denote survival functions from chemo and chemp+radio therapy groups, respectively
- *The goal is to find if $S_0(t)$ is 'different' from $S_1(t)$ using a suitable metric* (e.g. $D(a, b) = \int_a^b |S_0(t) - S_1(t)|dt$ 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 $T \perp C|Z$

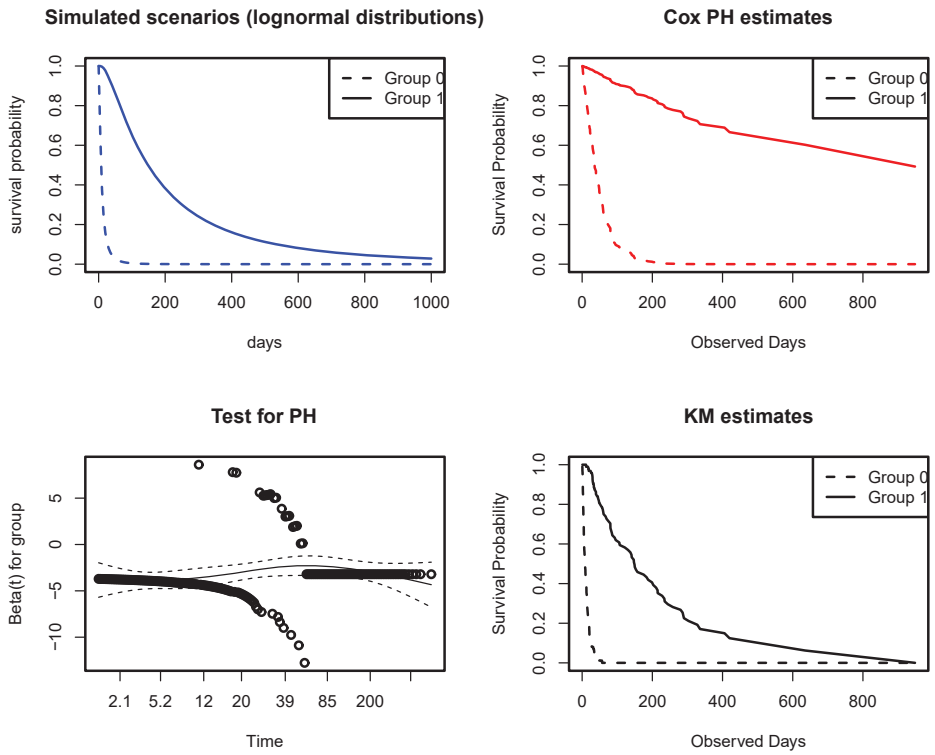
- How do we model the conditional survival function $S(t|z)$?
- Three popular models are based on conditional hazard function $h(t|z) = \frac{\partial}{\partial t}(-\log S(t|z))$. Let $h_j(t) = h(t|z = j)$ for $j = 0, 1$
  - (i) Proportional hazards (PH): $h_1(t) = \eta h_0(t)$ for some $\eta > 0$
    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 $\eta > 0$
- There are only two possibilities for any of these three models:
  - (a) If $\eta > 1$, then $S_1(t) < S_0(t)$ for all $t > 0$
  - (b) If $\eta < 1$, 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 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 $\eta = e^{-3}$)





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:

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

- We use the sequence of Bernstein basis functions: $g_{m,k}(t) = \frac{m}{\tau}\binom{m-1}{k-1}(\frac{t}{\tau})^{k-1}(1 - \frac{t}{\tau})^{m-k}\mathbb{I}(0 \le t \le \tau)$ for $m = 2, 3, \ldots$ which can be expressed as a density of $Beta(k, m - k + 1)$ at t/τ
- 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 $\int_0^\infty h(t)dt = \infty$ (because $S(\infty) = 0$)
- Thus, to complete the model specification, we define

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

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

$$\begin{aligned} l(\boldsymbol{\gamma}) &= \sum_{i=1}^{n} \{\Delta_i \log(h_m(Y_i, \boldsymbol{\gamma})) - H_m(Y_i, \boldsymbol{\gamma})\} \\ &= \sum_{i=1}^{n} \{\Delta_i \log(U_i^T \boldsymbol{\gamma}) - V_i^T \boldsymbol{\gamma}\}, \end{aligned} \tag{2}$$

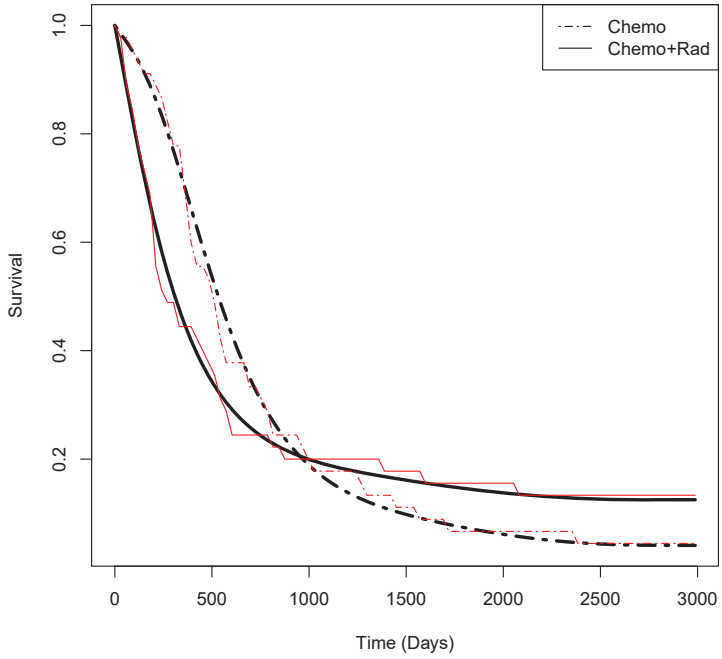
  where $\boldsymbol{\gamma} \in \mathcal{C}_m = [0, \infty)^m$, $U_i = \boldsymbol{g}_m(Y_i)$, and $V_i = \boldsymbol{G}_m(Y_i)$

- Notice that the existence and uniqueness of the (sieve) maximum likelihood estimator follows from strict concavity of the 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)
- It can be shown that *the form of log-likelihood of no-covariate case remains essentially the same of that with covariates*

Analysis of Gastric Cancer Data

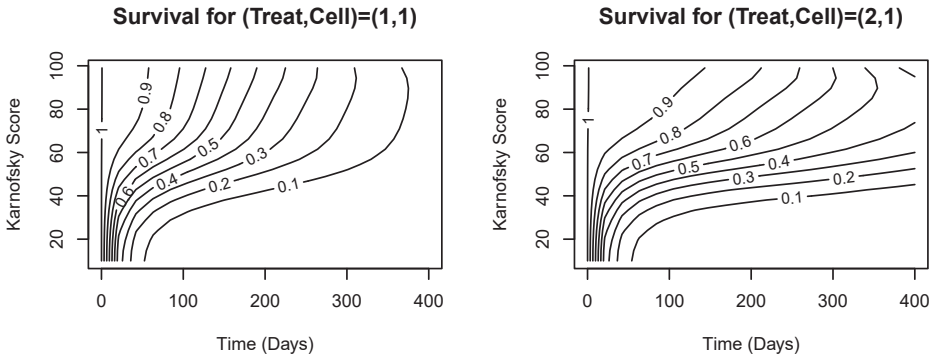
- Recall that $n = 90$ patients with locally advanced gastric carcinoma were randomized to two treatment groups (45 patients per group)
- One group only received chemotherapy while the other group received radiotherapy together with the same chemotherapy.
- As shown by the Kaplan-Meier curves, before the crossing point at approximately 1000 days the patients in the group receiving only chemotherapy 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 $m = [n^{0.5}] = 10$
- The results indicate that the estimated smooth curves cross at $t = 952$ days

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

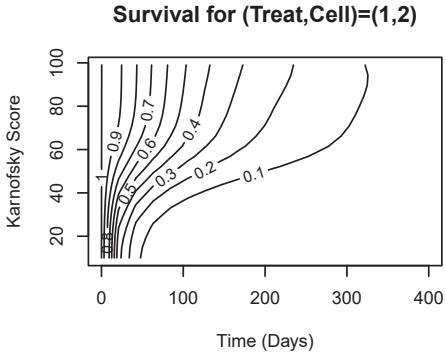


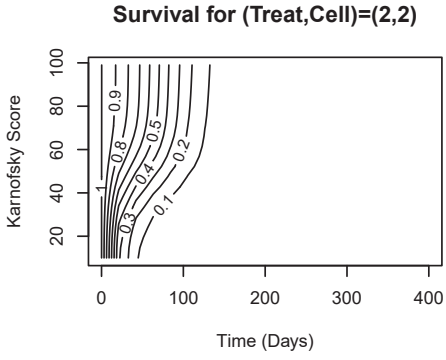
Analysis of Veterans Administration Data

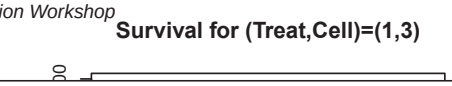
- In the Veterans Administration lung cancer study $n = 137$ male patients were assigned to two treatment groups: standard chemotherapy and test chemotherapy
- In addition, 5 other baseline covariates were recorded
- Following previous works we only used Karnofsky performance score and cell type as important 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 a partially linear coefficient model with the order $m = [n^{0.5}] = 12$

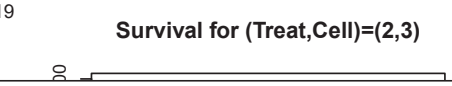

Survival for (Treat,Cell)=(1,1)
Karnofsky Score
Time (Days)
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1

Survival for (Treat,Cell)=(2,1)
Karnofsky Score
Time (Days)
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1

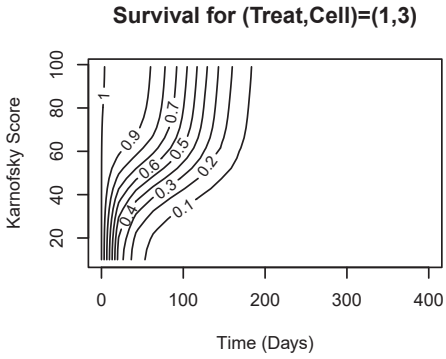

Survival for (Treat,Cell)=(1,2)
Karnofsky Score
Time (Days)
1
0.9
0.7
0.6
0.5
0.4
0.3
0.2
0.1

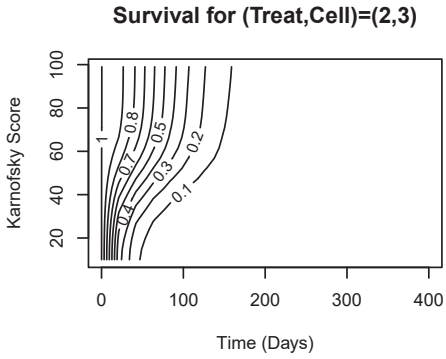

Survival for (Treat,Cell)=(2,2)
Karnofsky Score
Time (Days)
1
0.9
0.8
0.5
0.4
0.2
0.3
0.1

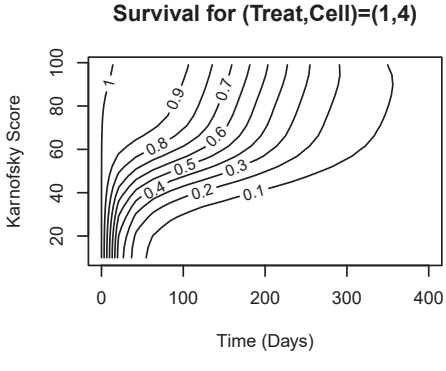

Survival for (Treat,Cell)=(1,3)
Time (Days)

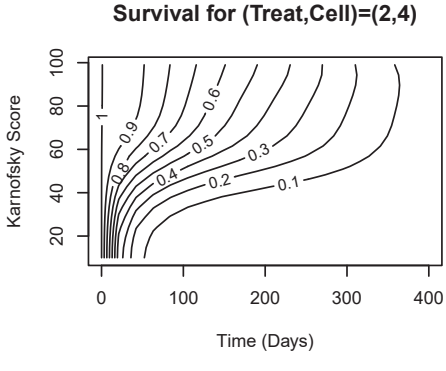

Survival for (Treat,Cell)=(2,3)
Time (Days)

(contd.)


Survival for (Treat,Cell)=(1,3)
Karnofsky Score
Time (Days)
1
0.9
0.7
0.6
0.5
0.4
0.3
0.2
0.1


Survival for (Treat,Cell)=(2,3)
Karnofsky Score
Time (Days)
1
0.8
0.7
0.5
0.3
0.2
0.4
0.1


Survival for (Treat,Cell)=(1,4)
Karnofsky Score
Time (Days)
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1


Survival for (Treat,Cell)=(2,4)
Karnofsky Score
Time (Days)
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1

- 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

- 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
- Data driven choice of m could deserve more future studies
- Extension of the proposed model to high-dimensional covariates requires careful analysis

Questions?

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

For R code email: sujit.ghosh@ncsu.edu

Theoretical Properties

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

$$d(\theta_1, \theta_2) = \left\{ \int (\sqrt{p_{\theta_1}} - \sqrt{p_{\theta_2}})^2 d\mu \right\}^{1/2}, \quad (3)$$

  where $\theta_j = \theta_j(\cdot)$ denote hazard functions and p_{θ_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

$$\Theta = \{\theta(\cdot) \in C[0,\tau] : \theta(\cdot) \text{ satisfies (I)-(III)}\}. \quad (4)$$

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

$$\Theta_m = \left\{ \theta_m(t) = \sum_{k=1}^{m} \gamma_k g_{m,k}(t) : \boldsymbol{\gamma} = (\gamma_1, \gamma_2, ..., \gamma_m)^T \in [0, L_\gamma]^m \right\}, \quad (5)$$

Theorem 1. *(Consistency) Suppose the conditions (I)-(II) hold and the sieve Θ_m is defined as in (5), then $\hat{h}_{m,n}(\cdot) \to_{a.s.} h_0(\cdot)$ 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 (5), 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)