Lectures of Machine learning in Survival analysis.

Hsieh Fushing *
Department of Statistics, University of California, Davis.

October 3, 2018

Abstract

In this lecture note, we study classic survival analysis, at the same time we challenge ourselves to build machine learning algorithm for survival analysis.

Lecture 1: From Cohort life table to Kaplan-Meier estimate.

Consider two contrasting settings:

[A](Complete data setting) : Let n subjects be taken from one designated population and observed with compete survival time $\{T_i\}$ for $i = 1, ...n$. Let T_i be I.I.D with respect to a distribution function $F(t) = 1 - S(t)$ for $t \in R^+ = [0, \infty)$. The classical empirical distribution function:

$$F_n(t) = \frac{1}{n}\sum_{i=1}^{n} 1_{[T_i \le t]},\ t \in R^+.$$

And the variance of $F_n(t)$ is calculated via Central Limit Theorem as

$$Var[F_n(t)] = Var[1 - F_n(t)] = Var[S_n(t)] = \frac{F(t)S(t)}{n}.$$

Here $S(t)$ is called survival function and $\hat{Pr}[T > t] = S_n(t)$.

[B](Compromised data setting): Let n subjects be also taken from one designated population. Here we also let T_i be I.I.D with respect to a distribution function $F(t) = 1 - S(t)$ for $t \in R^+ = [0, \infty)$, and L_i is i-th subject's potential lost-to-follow-up time, W_i is i-th subject's potential withdraw time. The compromised dataset contains data derived from ith subject in a form of observed compete survival time T_i, or simply one of the two binary status of L_i and W_i regarding to which interval they are falling into for $i = 1, ..., n$. It is note that compete data of

*Correspondence: Hsieh Fushing, University of California at Davis, CA, 95616. E-mail: fhsieh@ucdavis.edu

Table 1: *Cohort Life Table*

I_1	I_2	I_3	I_4	I_5	I_6	I_7	I_8	I_9	I_{10}	I_{11}
n_1	n_2	n_3	n_4	n_5	n_6	n_7	n_8	n_9	n_{10}	n_{11}
d_1	d_2	d_3	d_4	d_5	d_6	d_7	d_8	d_9	d_{10}	d_{11}
l_1	l_2	l_3	l_4	l_5	l_6	l_7	l_8	l_9	l_{10}	l_{11}
w_1	w_2	w_3	w_4	w_5	w_6	w_7	w_8	w_9	w_{10}	w_{11}

L_i and W_i are never observed. With such a compromised dataset, we can NOT estimate $F(t)$ or $S(t) = pr[T > t]$ here as easily as we can in setting A.

The difficulty facing setting B is that the I.I.D framework for $\{T_i\}$ has been compromised by being not able to see some T_is completely. This were one key problem facing the insurance company back to 200 years ago. The invention of Cohort-Life-Table is a technical advance. Its essence is to have the compromising effects of $\{L_i, W_i\}_1^n$ being separated and compartmented within suitable intervals. See an illustrating example with following 11 intervals as in Table 1. here d_k, l_k and w_k are counts of observed T_i, L_i and W_i falling in the interval $I_k = [\tau_{k-1}, \tau_k]$, and we have $n_{k+1} = n_k - d_k - l_k - w_k$ for all $k = 1, ..., K(= 11)$.

By making use of the effective base sample-size $n_k^* = n_k - (l_k + w_k)/2$, the insurance industry employed a product estimate for survival function $S(\tau_k)$

$$\begin{aligned} \hat{Pr}[T > \tau_k | T > \tau_{k-1}] &= \frac{n_k^* - d_k}{n_k^*}, \\ \hat{Pr}[T > \tau_k] = \hat{S}_n(\tau_k) &= \prod_{h=1}^{k} \hat{Pr}[T > \tau_h | T > \tau_{h-1}]. \end{aligned}$$

To evaluate the $Var[\hat{Pr}[T > \tau_k]]$, we proceed by taking the logarithm on

$$\log \hat{Pr}[T > \tau_k] = \sum_{h=1}^{k} \log \hat{Pr}[T > \tau_h | T > \tau_{h-1}],$$

and by $\delta-$method we have

$$\begin{aligned} & var[\log \hat{Pr}[T > \tau_h | T > \tau_{h-1}]] \\ = & \frac{1}{\hat{Pr}^2[T > \tau_h | T > \tau_{h-1}]} \frac{\hat{Pr}[T > \tau_h | T > \tau_{h-1}](1 - \hat{Pr}[T > \tau_h | T > \tau_{h-1}])}{n_h^*} \\ = & \frac{d_h}{n_h^*(n_h^* - d_h)}. \end{aligned}$$

Then a big assumption is taken to claiming "some sort of independence" among $\{\hat{Pr}[T > \tau_h | T > \tau_{h-1}]\}$, such that

$$Var[\log \hat{Pr}[T > \tau_k]] = \sum_{h=1}^{k} var[\log \hat{Pr}[T > \tau_h | T > \tau_{h-1}]].$$

Table 2: *Cohort Life Table* under setting (C)

I_1	I_2	I_3	I_4	I_5	I_6	I_7	I_8	I_9	I_{10}	I_{11}
n_1	n_2	n_3	n_4	n_5	n_6	n_7	n_8	n_9	n_{10}	n_{11}
δ_1	δ_2	δ_3	δ_4	δ_5	δ_6	δ_7	δ_8	δ_9	δ_{10}	δ_{11}
$1-\delta_1$	1-δ_2	1-δ_3	1-δ_4	1-δ_5	1-δ_6	1-δ_7	1-δ_8	1-δ_9	1-δ_{10}	1-δ_{11}

This "claim" turns out to be asymptotically "correct".

Again by applying the $\delta-$ method we can derive the Greenwood formula:

$$Var[\hat{Pr}[T > \tau_k]] \approx \hat{S}_n^2(\tau_k) \sum_{h=1}^{k} \frac{d_h}{n_h^*(n_h^* - d_h)}$$

[C](Right censoring) When L_i and W_i are in fact observable, we take them as equal and denote it as C_i, which then call right censoring variable. The data format become $\{(X_i, \delta_i)|X_i = T_i \wedge C_i, \delta_i = 1_{[T_i \leq C_i], i=1,,n}\}$. That is, δ_i is the censoring status. The question remains: How to estimate $S(t)$ for all $t \in R^+$?

Let the order statistics of $\{X_i\}_1^n$ be denoted as $\{X_{(i)}\}_1^n$. For the time being, let's assume no ties. Then we take $X_{(k)} = \tau_k$ for all $k = 1, n(= K)$. We then construct a "life table" as in Table 2:

Then we have an estimate of $S(t)$ as

$$\begin{aligned} \hat{S}_{KM}(t) = \prod_{X_{(i)} \leq t} (1 - \frac{\delta_{(i)}}{n - i + 1 - \frac{1-\delta_{(i)}}{2}} &= \prod_{X_{(i)} \leq t} (1 - \frac{\delta_{(i)}}{n - i + 1}) \\ &= \prod_{X_{(i)} \leq t} (\frac{n-i}{n-i+1})^{\delta_{(i)}}. \end{aligned}$$

This is so called Kaplan-Meier estimate of $S(t)$. Further, via Green wood formula, we have

$$Var[\hat{S}_{KM}(t)] = (\hat{S}_{KM}(t))^2 \sum_{X_{(i)} \leq t} \frac{\delta_{(i)}}{(n-i)(n-i+1)}.$$

Homework # 1 First, you transform your dataset into a cohort life table and then estimate all $\{S(\tau_k)\}_1^K$ and their variances via Greenwood formula. Secondly calculate the $\hat{S}_{KM}(t)$ and compare its variances with that of $\hat{S}(\tau_k)$ at all chosen $\{\tau_k\}$. (Note: if your dataset contains covariate information, like $\{(X_i, Z_i, \delta_i)\}$, then ignore the Z_i for the time being.)

Homework # 2 How to mimic the data set $\{(X_i, \delta_i)$? (Hit: Consider the case of complete data, first.)

References

[1] Hsieh Fushing. and Roy T. Complexity of Possibly-gapped Histogram and Analysis of Histogram (ANOHT). *Royal Socity-Open Science*, 2018.

[2] Hsieh Fushing, Liu S.-Y., Hsieh Y.-C., and McCowan B. From patterned response dependency to structured covariate dependency: categorical-pattern-matching. *PLoS One*,2018.

Lectures of Machine learning in Survival analysis.

Hsieh Fushing *
Department of Statistics, University of California, Davis.

October 9, 2018

Lecture 2: From Kaplan-Meier estimate to Nelson-Aallen estimates.

0.1 Redistribution-to-the-right

Consider a set of possibly right censored $\{(X_i, \delta_i)\}_1^n$, and denote its ordered statistics as $\{(X_{(i)}, \delta_{(i)})\}_1^n$. The Kaplan-Meier estimate is:

$$\hat{S}_{KM}(t) = \prod_{X_{(i)} \leq t} (1 - \frac{\delta_{(i)}}{n-i+1}).$$

It is known that $\hat{S}_{KM}(t)$ only have jumps at uncensored time points, that is, for those is with $\delta_i = 1$. What are the sizes at all jumps? It is noted that the sizes are constant $1/n$ when there is no censoring involved. For illustration purpose, let's assume that $\delta_{(1)} = \delta_{(2)} = \delta_{(4)} = \delta_{(6)} = 1$ and $\delta_{(3)} = 0$. Let's define $X_{(t_i)} \leq t_i < X_{(i+1)}$ for all is. Then we have

$$\begin{aligned}
\hat{S}_{KM}(t_1) &= 1 - \frac{1}{n}; \\
\hat{S}_{KM}(t_2) &= 1 - \frac{2}{n} = \hat{S}_{KM}(t_3) \\
\hat{S}_{KM}(t_4) &= \hat{S}_{KM}(t_2)(1 - \frac{1}{n-4+1}) = \hat{S}_{KM}(t_2) - (1 - \frac{2}{n})\frac{1}{n-3} = \hat{S}_{KM}(t_2) - \frac{1}{n}(1 + \frac{1}{n-3}).
\end{aligned}$$

Therefore we have the jump sizes calculated as

$$\begin{aligned}
\Delta\hat{S}_{KM}(X_{(1)}) &= \frac{1}{n} = \Delta\hat{S}_{KM}(X_{(2)}); \\
\Delta\hat{S}_{KM}(X_{(3)}) &= 0; \\
\Delta\hat{S}_{KM}(X_{(4)}) &= \frac{1}{n} + \frac{\frac{1}{n}}{n-3}.
\end{aligned}$$

Further we can calculated the following:

*Correspondence: Hsieh Fushing, University of California at Davis, CA, 95616. E-mail: fhsieh@ucdavis.edu

1 If $\delta_{(5)} = 1$, then we have

$$\Delta\hat{S}_{KM}(X_{(5)} = \Delta\hat{S}_{KM}(X_{(4)} = \Delta\hat{S}_{KM}(X_{(6)}.$$

2 If $\delta_{(5)} = 0$, then we have

$$\begin{aligned} \Delta\hat{S}_{KM}(X_{(5)}) &= 0; \\ \Delta\hat{S}_{KM}(X_{(6)} &= \frac{1}{n} + \frac{\frac{1}{n}}{n-3} + \frac{\frac{1}{n} + \frac{1}{n(n-3)}}{n-5}. \end{aligned}$$

This phenomenon is called re-distribution-to-the right, which was discovered by B. Efron (1967).

The discrete jump sizes at all uncensored time points allow us to think about one important question: How to build a histogram based on $\{(X_i, \delta_i)\}_1^n$? We explore two versions of such constructions for histograms based on $\hat{S}_{KM}(t)$.

Mimicking through $\hat{S}_{KM}(t)$: v-1 : Apply Hierarchical Clustering (HC) algorithm on uncensored data points $\{X_i|\delta_i = 1\}^{n_u}$, and then adjust weights in each bin according to the total jump sizes contained.

Mimicking through $\hat{S}_{KM}(t)$: v-2 : We build the following algorithm:

1. Define
$$s_{(i)} = \frac{\Delta\hat{S}_{KM}(X_{(i)})}{\Delta X_{(i)}}$$
Where
$$\Delta X_{(i)} = X_{(i)} - \max\{X_{(j)}|j < i,\ \delta_{(j)} = 1\}$$
for all uncensored time points.

2 Apply Hierarchical Clustering (HC) algorithm on $\{s_{(i)}\}^{n_u}$, and select a tree level for a clustering composition, in which each cluster is called a slope-state.

3 Layout the slope-states upon the all locations of $X_{(i)}$, where $\delta_{(i)} = 1$.

4 Mark the "run" of each slope-state. Here a "run" is a consecutively repeating slope-state segment.

5 Make each run into one bin.

The reason behind constructing histograms is that a histogram provide a proper platform for mimicking original data set $\{(X_i = T_i \wedge C_i, \delta_i)\}_1^n$.

0.2 Nelson-Aalen estimate

If we evaluate the risk of an event w.r.t a continuous event time T through the conditional probability:

$$Pr[T < t + \Delta | T > t] = \frac{Pr[t < T < t + \Delta]}{Pr[T > t]} \approx \frac{F'(t)}{S(t)}\Delta = \lambda_T(t)\Delta.$$

The risk evaluation via hazard function $\lambda_T(t)$ has many desired properties:

$$\begin{aligned} \Lambda(t) &= \int_0^t \lambda_T(s)ds = \int_0^t \frac{F'(s)}{S(s)}ds = -lnS(t) \\ S(t) &= e^{-\Lambda(t)}. \end{aligned}$$

The cumulative hazard function $\Lambda(t)$ can be easily estimated via $\hat{S}_{KM}(t)$ as

$$\hat{\Lambda}(t) = -ln\hat{S}_{KM}(t) = -\sum_{X_{(i)} \le t} \log(1 - \frac{\delta_{(i)}}{n - i + 1})$$

$$\hat{\Lambda}_{NA}(t) = \sum_{X_{(i)} \le t} \frac{\delta_{(i)}}{n - i + 1}.$$

By applying $\delta-$ method, we have a simple formula for variance of $\hat{\Lambda}_{NA}(t)$,

$$Var[\hat{\Lambda}_{NA}(t)] = \sum_{X_{(i)} \le t} \frac{\delta_{(i)}}{(n - i + 1)(n - i)}.$$

This formula can be proved via Martingale central limit theorem in counting process theory. The simplicity, seen through through the summation in the above formula, implies a very concise convergence theory pertaining to $\hat{\Lambda}_{NA}(t)$.

Mimicking through $\hat{\Lambda}_{NA}(t)$: We build the following algorithm:

1. Define

$$l_{(i)} = \frac{\hat{\Lambda}_{NA}(X_{(i)})}{\Delta X_{(i)}}$$

Where

$$\Delta X_{(i)} = X_{(i)} - \max\{X_{(j)} | j < i,\ \delta_{(j)} = 1\}$$

for all uncensored time points.

2 Apply Hierarchical Clustering (HC) algorithm on $\{l_{(i)}\}^{n_u}$, and select a tree level for a clustering composition, in which each cluster is called a slope-state.

3 Layout the slope-states upon the all locations of $X_{(i)}$, where $\delta_{(i)} = 1$.

4 Mark the "run" of each slope-state. Here a "run" is a consecutively repeating slope-state segment.

5 Make each run into one bin.

It is noted that a piece-wise constant segment on $\hat{\Lambda}_{NA}(t)$ means an exponential component, $\exp(\lambda)$, in the random mechanism, which generates the observed data.

Homework # 2 extended How to mimic the data set $\{(X_i = T_i \wedge C_i, \delta_i)$? (Hit: Consider the case of complete data, first and duality of roles T_i and C_i.)? Compare the 3 mimicking algorithms.

R-package from Aleksandra Taranov $< ataranov@ucdavis.edu >$ **without censoring setting.** These are the instructions to open the R package (to be run from within Rstudio).

#To install within Rstudio, open Rstudio and in it, run:
install.packages("devtools")
$devtools :: install_github("taranov2007/GappedHist")$
# To test out the functions, run:
library(GappedHist)
$X < -iris[, 1:4]$
head(X)
$M3 = histbyDESS(values = X[, 3], epsilon = 1, graph = TRUE); M3$
$par(mfrow = c(1, 3))$
$cdffromtable(X[, 3], M3\$table)$
$histfromtable(M3\$table)$
$plotDESS(X[, 3], M3)$

References

[1] Hsieh Fushing. and Roy T. Complexity of Possibly-gapped Histogram and Analysis of Histogram (ANOHT). *Royal Socity-Open Science*, 2018.

[2] Hsieh Fushing, Liu S.-Y., Hsieh Y.-C., and McCowan B. From patterned response dependency to structured covariate dependency: categorical-pattern-matching. *PLoS One*,2018.

Lectures of Machine learning in Survival analysis.

Hsieh Fushing *
Department of Statistics, University of California, Davis.

October 16, 2018

1 Lecture 3: Multiple sample Problems with right censored data.

1.1 Analysis of Histogram(ANOHT)

Let's consider the K-sample problem with right censored data: $\{(X_i^k, \delta_i^k)|i =, 1, ..., n_k; k = 1, ..., , K\}$ with $X_i^k = T_i^k \wedge C_i^k$ and T_i^k and C_i^k being stochastically independent. Denote that T_i^k is distributed according to survival and cumulative hazard functions as $S^k(t) = e^{-\Lambda^k(t)}$.

Question of interest : Where are the temporal localities of difference among K survival functions $\{S^k(t)\}_1^K$ and cumulative hazard functions $\{\Lambda^k(t)\}_1^K$?

Algorithm for K-sample Problem : We develop the following algorithm to discover potential temporal localities of differences of $\{S^k(t)\}_1^K$ and $\{\Lambda^k(t)\}_1^K$. Encode each sample with a color, and denote the overall null Shannon entropy as

$$\xi[null] = -\sum_1^K \frac{n_k}{N} \log \frac{n_k}{N}$$

Where $N = n_1 + ... + n_K$.

[KSP-1] We pool all data points from the K samples into one $\{(Z_j, \eta_j)|j =, 1, ..., N;\}$ and construct one version of histogram either based on Kaplan-Meier or Nelson Aalen estimates as discussed in the Lecture-notes 2 or see [1];

[KSP-2] Upon each bin, say $(a_h, b_h]$, we recover the color-code proportions as $\{p^k[h]\}_1^K$ and compute the bin-specific Shannon entropy as

$$\xi[h]_0 = -\sum_1^K p^k[h] \log\{p^k[h]\}.$$

*Correspondence: Hsieh Fushing, University of California at Davis, CA, 95616. E-mail: fhsieh@ucdavis.edu

[KSP-3] Mimic the K-sample $\{(X_i^{k*}, \delta_i^{k*})|i =, 1, .., n_k; k = 1, ..., , K\}$ for M times;

[KSP-4] For $m-$th mimicry of K-sample, we calculate the color-code proportions as $\{p^{k*}[h]_m\}_1^K$ upon all bins $\{(a_h, b_h]\}_1^H$ and calculate the Shannon entropy as:

$$\xi^*[h]_m = -\sum_1^K p^{k*}[h]_m \log\{p^{k*}[h]_m\};$$

[KSP-4] Draw an empirical distribution $\hat{D}[h](.)$ based on $\{\xi^*[h]_m\}_1^M$ and compare it with $\xi[h]_0$ as well as $\xi[null]$ by evaluating $\hat{D}[h](\xi[h]_0)$ and $\hat{D}[h](\xi[null])$, respectively.

If the value $\hat{D}[h](\xi[null])$ is extreme large as being very close to 1 for any bin $(a_h, b_h]$, that is, almost of all $\{\xi^*[h]_m\}_1^M$ are smaller than $\hat{D}[h](\xi[null])$. Then we can claim we discover a tempo-locality where $\{S^k(t)\}_1^K$ or $\{\Lambda^k(t)\}_1^K$ are unequal. In fact we can see which colors are over-representing and which are under-representing.

As a remark, if the independent censoring assumption is correct, then we expect that $\hat{D}[h](\xi[h]_0)$ is around $\frac{1}{2}$. Otherwise, the assumption might possibly fail at some samples.

We then address how to test the classic hypothesis: H_0 : all survival functions $\{S^k(t)\}_1^K$, or cumulative functions $\{\Lambda^k(t)\}_1^K$, are equal. We can synthesize above pieces of evidence into one. Let $\{w_h\}$ be the weights of bin found on $(a_h, b_h]$ in the Step-[SKP-1]. We calculate an overall directed conditional entropy, see also [2], as:

$$\xi[OA] = \sum_{h=1}^H w_h \xi[h]_0$$

And likewise for each mimicry of K-sample data: for all m=1, ,,M,

$$\xi[OA]_m = \sum_{h=1}^H w_h \xi^*[h]_m.$$

We then build an empirical distribution $\hat{D}[OA](.)$ and calculate the value $\hat{D}[OA](\xi[null])$. We can make use of $1 - \hat{D}[OA](\xi[null])$ as a p-value in statistical testing hypothesis.

1.2 Simultaneous Multiple Comparison

With the K-sample data, we are also interested in the following Simultaneous Multiple Comparison question, which is beyond classical pairwise comparison, such as Tukey's Simultaneous Multiple Comparison question.

Simultaneous Multiple Comparison question : Which samples are more similar with each other, but much less similar with which samples? We develop an algorithm to resolve the question by discovering a tree on many mimicries of the K samples. This tree offers us a global view of the K-samples as well as local views among different samples.

Algorithm for multiple comparison on K-sample Problem :

[KSP-MC-1] Construct) mimicries of each sample of $\{(X_i^{k*}, \delta_i^{k*})|i = 1, ..., n_k; k = 1, ..., , K\}$;

[KSP-MC-2] Transform each mimicry into a $1 \times H$-vector of weights, say $\mathcal{W}_m^k = (w^k[1]_m, ..., w^k[H]_m)$, calculated as increments of Kaplan-Meier or Nelson-Aalen estimates across all H bins $\{(a_h, b_h]\}_1^H$;

[KSP-MC-3] Stacking all row vectors $\{\mathcal{W}_m^k|k = 1, ..., K; m = 1, ..., M\}$ into a $(K \times M) \times H$ matrix;

[KSP-MC-4] Build a HC tree, say $\mathcal{T}^K[M]$, on the row axis.

This tree $\mathcal{T}^K[M]$ will allow us to see which groups of samples are far away from which groups by being located at different big branches. This is the global view. For local view, we see what degrees of mixing among mimicries of samples to determine how close they are.

1.3 Gaussian process theory

A Gaussian process $Z(t)$ for $t \in R^+$ is defined to have the following to properties:

G1: $Z(t)$ is distributed with respect to $N(\mu(t), \sigma(t, t))$;

G2: For any temporal m-vector $(t_1, t_2, , t_m)$ with $t_i < t_{i+1}$ for all $i < m$, the m-vector $(Z(t_1), Z(t_2), , Z(t_m))$ is multivariate normal distributed according to $N(\mu, \Sigma_m)$ with mean vector $(\mu(t_1), \mu(t_2), , \mu(t_m))$ and $m \times m$ covariance matrix $\Sigma_m = [\sigma(t_i, t_j)]$ for $1 \leq i, j \leq m$.

Here the covariance function $\sigma(s, t) = Cov[Z(s), Z(t)]$ for $s, t \in R^+$. A Gaussian process $Z(t)$ is centering around 0 if $\mu(t) \equiv 0$, so such a Gaussian process is characterized by its covariance function $\sigma(s, t)$.

There are two well-known special cases: Brownian motion $W(t)$ for $t \in R^+$ and Brownian bridge $B(t)$ $t \in [0.1]$.

Brownian motion $W(t)$

1: $W(t) \sim N(0, t)$;

2: $\sigma(s, t) = s \wedge t = \min(s, t)$.

The key property of Brownian motion is called "independent increment property", that is, $W(t) \perp W(t + s) - W(t)$. The random increment $W(t + s) - W(t) \sim N(0, s)$ is stochastically independent of $W(t)$, where the increment begins. This is an important property.

Brownian Bridge $B(t)$. For all $t \in [0, 1]$

$$B(t) = W(t) - tW(1).$$

Therefore we have, if $s < t$, then

1: $B(t) \sim N(0, t(1 - t))$;

2: $\sigma(s,t) = s(1-t)$.

One well-known example of Brownian Bridge process is the convergent Gaussian process of $\sqrt{n}(F_n(t)-F(t))$ under i.i.d setting with complete data: as $n \to \infty$, the $F_n(t) = 1/n \sum_{i=1}^{n} 1_{[X_i \le t]}$,

$$\sqrt{n}(F_n(t) - F(t)) \to B(F(t)).$$

Asymptotic theory on Nelson-Aalen estimation One very important asymptotical result in Survival Analysis is related to convergence of Nelson-Aalen estimate: under the right censoring setting with data denoted as $\{(X_i(= T_i \wedge C_i), \delta_i)\}$ with $X_i = T_i \wedge C_i$ and $T_i \perp C_i$, and $T_i \sim F(t)(= 1 - S(t))$ and $C_i \sim G(t)$,

$$\sqrt{n}(\Lambda_{NA}(t) - \Lambda(t)) \to Z(t),$$

where $Z(t)$ is a Gaussian process centering around 0, and has a special covariance function with independent increment property as:

$$\sigma(s,t) = \int_0^{s \wedge t} \frac{dF_u(y)}{[1-H(y)]^2}$$

with

$$\begin{aligned} F_u(t) &= Pr[X_i < t, \delta_i = 1] \\ &= Pr[T_i < C_i, T_I < t] = \int_0^t (\int_s^\infty g(c)dc) f(s)ds \\ &= \int_0^t (1 - G(c)) f(s)ds, \end{aligned}$$

and

$$\begin{aligned} H(t) &= Pr[X_i < t] \\ 1 - H(t) &= (1 - F(t))(1 - G(t)) \end{aligned}$$

It is also essential to note the following approximation

$$\hat{Cov}[Z(t), Z(t)] = \hat{\sigma}(t,t) = n \times \sum_{X_{(i)} \le t} \frac{\delta_{(i)}}{(n-i+1)(n-i)} \approx n \times Var[\hat{\Lambda}_{NA}(t)].$$

Therefore we have that, for any temporal m-vector $(t_1, t_2, , t_m)$ with $t_i < t_{i+1}$ for all $i < m$, the components of the m-vector $(Z(t_1), Z(t_2) - Z(t_1), , Z(t_i) - Z(t_{i-1}), , Z(t_m) - Z(t_{m-1}))$ are mutually independent and normal distributed according to $N(0, \sigma(t_i, t_i) - \sigma(t_{i-1}, t_{i-1}))$.

Asymptotic theory on Kaplan-Meier estimation Under the same right censoring setting, we have the following asymptotical result on Kaplan-Meier estimate of survival function $S(t)$ as

$$\sqrt{n}(\hat{S}_{KM}(t) - S(t)) \to Z_S(t),$$

where $Z_S(t)$ is a Gaussian process centering around 0, and has a special covariance function as:

$$\sigma_S(s,t) = S(s)S(t) \int_0^{s \wedge t} \frac{dF_u(y)}{[1-H(y)]^2}.$$

Hence it is noted that $Z_S(t)$ doesn't have the independent increment property.

Homework#3. For all students, perform the K-sample problems on your dataset by using the machine learning techniques presented in the Section 1 and 2 of this Lecture notes.

For PhD students, also perform the K-sample problems by adopting the asymptotical results presented in section 3 of this Lecture-notes, and then compare with your previous results. (Hint: Reference [1] is relevant.)

References

[1] Hsieh Fushing. and Roy T. Complexity of Possibly-gapped Histogram and Analysis of Histogram (ANOHT). *Royal Socity-Open Science*, 2018.

[2] Hsieh Fushing, Liu S.-Y., Hsieh Y.-C., and McCowan B. From patterned response dependency to structured covariate dependency: categorical-pattern-matching. *PLoS One*,2018.

Lectures of Machine learning in Survival analysis.

Hsieh Fushing *
Department of Statistics, University of California, Davis.

October 22, 2018

1 Lecture 4: Survival Analysis at the crossroad: I. nonparametric mimicking vs. parametric modeling on one sample problem.

1.1 Non-informative censoring?

In Survival analysis, upon a set of possibly right censored $\{(X_i, \delta_i)\}_1^n$ with with $X_i = T_i \wedge C_i$, we commonly assume the non-informative censoring assumption: $T_i \perp C_i$ being stochastically independent. This assumption indeed can be effectively checked by mimicking and machine learning techniques. The important question is: **what can we do and can't do if the assumption is rejected?**

Homework# 4-1. Check the non-informative censoring assumption in your data set.

1.2 Mean and median estimations

Assume the non-informative censoring assumption in process of generating the data set. Denote that T_i is distributed according to survival and cumulative hazard functions as $S(t) = e^{-\Lambda(t)}$, and denote its ordered statistics as $\{(X_{(i)}, \delta_{(i)})\}_1^n$. We consider the issues of estimating the mean $E[T_i] = \mu$ and its median $\theta = S^{-1}(1/2)$.

Estimation of $E[T_i] = \mu$. Estimation of mean under the i.i.d setting is easy because of Law of Large number and Central Limit Theorem. But it is not the straight forward under the right censoring setting because of uneven weighting among the uncensored data points and zero weighting on all censored data points. There are two approaches to perform mean estimation. The first one is given as follows.

*Correspondence: Hsieh Fushing, University of California at Davis, CA, 95616. E-mail: fhsieh@ucdavis.edu

If $t(S(t) \to 0$ as $t \to \infty$, via integration by part, then we have

$$\begin{aligned} E[T_i] &= \mu = \int_0^\infty t f(t) dt = -\int_0^\infty t dS(t) \\ &= (-1)[t(S(t)|_0^\infty - \int_0^\infty S(t) dt \\ &= \int_0^\infty S(t) dt. \end{aligned}$$

Therefore we can perform the following estimation: choose a set of H time points $0 = t_0 < t_1 < t_2 < t_h < t_H$ and denote $\Delta_h t = t_h - t_{h-1}$

$$\begin{aligned} \hat{\mu} &= \int_0^\infty \hat{S}_{KM}(t) dt \\ &\approx \sum_{h=1}^{H} \hat{S}_{KM}(t_h^*) \Delta_h t. \end{aligned}$$

Therefore we have:

$$\begin{aligned} \sqrt{n}(\hat{\mu} - \mu) &= \int_0^\infty \sqrt{n}(\hat{S}_{KM}(t) - S(t)) dt \\ &\approx \sum_{h=1}^{H} Z_{KM}(t_h^*) \Delta_h t. \end{aligned}$$

And by denoting

$$\begin{aligned} \widetilde{\Delta} t &= (\Delta_1 t, \Delta_2 t, , \Delta_h t, , \Delta_H t)^T \\ \widetilde{Z_{KM}} &= (Z_{KM}(t_1), Z_{KM}(t_2), , Z_{KM}(t_h), , Z_{KM}(t_H))^T, \end{aligned}$$

we also have the following

$$\begin{aligned} \widetilde{Z_{KM}}^T \widetilde{\Delta} t &= \sum_{h=1}^{H} \hat{S}_{KM}(t_h) \Delta_h t \sim N(\tilde{0}, (\widetilde{\Delta} t)^T \Sigma_{KM}[H] \widetilde{\Delta} t). \\ \Sigma_{KM}[H] &= [Cov(Z_{KM}(t_i), Z_{KM}(t_j))]_{H \times H} = [S(t_i) S(t_j) \int_0^{t_i \wedge t_j} \frac{dF_u(y)}{[1 - H(y)]^2}]_{H \times H} \end{aligned}$$

with

$$\Sigma_{KM}[H] = \begin{bmatrix} S(t_1) & 0 & 0 & \dots & 0 \\ 0 & S(t_2) & 0 & \dots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \dots & S(t_H) \end{bmatrix} \Sigma[H] \begin{bmatrix} S(t_1) & 0 & 0 & \dots & 0 \\ 0 & S(t_2) & 0 & \dots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \dots & S(t_H) \end{bmatrix}$$

and

$$\Sigma[H] = \begin{bmatrix} \int_0^{t_1\wedge t_1} \frac{dF_u(y)}{[1-H(y)]^2} & \int_0^{t_1\wedge t_2} \frac{dF_u(y)}{[1-H(y)]^2} & \int_0^{t_1\wedge t_3} \frac{dF_u(y)}{[1-H(y)]^2} & \cdots & \int_0^{t_1\wedge t_H} \frac{dF_u(y)}{[1-H(y)]^2} \\ \int_0^{t_1} \frac{dF_u(y)}{[1-H(y)]^2} & \int_0^{t_2} \frac{dF_u(y)}{[1-H(y)]^2} & \int_0^{t_2} \frac{dF_u(y)}{[1-H(y)]^2} & \cdots & \int_0^{t_2} \frac{dF_u(y)}{[1-H(y)]^2} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \int_0^{t_1} \frac{dF_u(y)}{[1-H(y)]^2} & \int_0^{t_2} \frac{dF_u(y)}{[1-H(y)]^2} & \int_0^{t_3} \frac{dF_u(y)}{[1-H(y)]^2} & \cdots & \int_0^{t_H} \frac{dF_u(y)}{[1-H(y)]^2} \end{bmatrix}$$

$$= \begin{bmatrix} 1 & 0 & 0 & \dots & 0 \\ 1 & 1 & 0 & \dots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & 1 & 1 & \dots & 1 \end{bmatrix} \begin{bmatrix} \int_0^{t_1} \frac{dF_u(y)}{[1-H(y)]^2} & 0 & 0 & \dots & 0 \\ 0 & \int_{t_1}^{t_2} \frac{dF_u(y)}{[1-H(y)]^2} & 0 & \dots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \dots & \int_{t_{H-1}}^{t_H} \frac{dF_u(y)}{[1-H(y)]^2} \end{bmatrix} \begin{bmatrix} 1 & 1 & 1 & \dots & 1 \\ 0 & 1 & 1 & \dots & 1 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \dots & 1 \end{bmatrix}$$

Therefore we have

$$\begin{aligned} (\widetilde{\Delta}t)^T \Sigma_{KM}[H] \widetilde{\Delta}t &= (\sum_{h=1}^H S(t_h)\Delta_h t, \sum_{h=2}^H S(t_h)\Delta_h t, \sum_{h=3}^H S(t_h)\Delta_h t, \dots, S(t_H)\Delta_H t) \times \\ &\quad \begin{bmatrix} \int_0^{t_1} \frac{dF_u(y)}{[1-H(y)]^2} & 0 & 0 & \dots & 0 \\ 0 & \int_{t_1}^{t_2} \frac{dF_u(y)}{[1-H(y)]^2} & 0 & \dots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \dots & \int_{t_{H-1}}^{t_H} \frac{dF_u(y)}{[1-H(y)]^2} \end{bmatrix} \begin{bmatrix} \sum_{h=1}^H S(t_h)\Delta_h t \\ \sum_{h=2}^H S(t_h)\Delta_h t \\ \sum_{h=3}^H S(t_h)\Delta_h t \\ \dots \\ S(t_H)\Delta_H t \end{bmatrix} \\ &= \sum_{j=1}^H \{\sum_{h=j}^H S(t_h)\Delta_h t\}^2 \int_{t_{h-1}}^{t_h} \frac{dF_u(y)}{[1-H(y)]^2} \\ &\approx \int_0^\infty \frac{dF_u(y)}{[1-H(y)]^2} (\int_y^\infty S(t)dt)^2. \end{aligned}$$

Homework# 4-2. Derive the mean estimate using the histogram and its variation based on mimicking. PhD students should compare histogram based results with the estimating approach based on integration and its approximated variance.

Estimation of median $\theta = S^{-1}(1/2)$. Let's consider $\hat{\theta}$ such that $\hat{S}_{KM}(\hat{\theta}) = 1/2$. Then we can find

$$\hat{\theta} = \inf\{t | \hat{S}_{KM}(t) \le 1/2\}.$$

Denote the true parameter as $\theta_0 = S^{-1}(1/2)$. Then we have:

$$0 = \hat{S}_{KM}(\hat{\theta}) - S(\theta_0) = \hat{S}_{KM}(\hat{\theta}) - S(\hat{\theta}) + S(\hat{\theta}) - S(\theta_0)$$

If $\hat{\theta}$ is somehow close to θ_0, then we more or less have:

$$\hat{S}_{KM}(\hat{\theta}) - S(\hat{\theta}) \approx \hat{S}_{KM}(\theta_0) - S(\theta_0)$$

and

$$S(\hat{\theta}) - S(\theta_0) \approx S'(\theta_0)(\hat{\theta}) - \theta_0).$$

Therefore we arrive at

$$\begin{aligned}
\sqrt{n}(\hat{\theta}) - \theta_0) &\approx \frac{\sqrt{n}(\hat{S}_{KM}(\theta_0) - S(\theta_0))}{S'(\theta_0)}, \\
&\approx \frac{Z_{KM}(\theta_0)}{S'(\theta_0)} \\
&\sim N(0, \frac{\int_0^{\theta_0} \frac{dF_u(y)}{[1-H(y)]^2}}{4f^2(\theta_0)}).
\end{aligned}$$

1.3 Parametric estimations

The most fundamental distribution in survival analysis is the Exponential distribution with a constant hazard function λ, denoted as $EXP(\lambda)$. This constancy in hazard is termed memory-less property: A random variable, say T, always forgets how long it has been in waiting for the event to come:

$$E[T] = E[T|T > t_0].$$

The cumulative hazard function and survival function are given as follows:

$$\Lambda(t) = \lambda t = -\log S(t).$$

With $a > 0$, we have its exponential $W = T^a$ with a survival function as

$$S_W(t) = Pr[T^a > t] = Pr[T > t^{1/a}] = S_T(t^{1/a}) = e^{-\lambda t^{1/a}}.$$

Denote $1/a = \alpha$, we have

$$\Lambda_W(t) = \lambda t^\alpha = \int_0^t \lambda \alpha s^{\alpha-1} ds.$$

Such a W is call Weibull distribution $W(\lambda, \alpha)$.

Consider what a data point (X_i, δ_i) contributes to the likelihood function.

1. if $\delta_i = 1$, this compete data point contribute

$$f(X_i|(\lambda, \alpha)) = \lambda \alpha X_i^{\alpha-1} e^{-\lambda X_i^\alpha};$$

2. if $\delta_i = 0$, this incomplete data point contribute

$$S(X_i|(\lambda, \alpha)) = e^{-\lambda X_i^\alpha}.$$

Therefore, given $\{(X_i, \delta_i)\}_1^n$, the likelihood function of (λ, α) is

$$\begin{aligned} L_n(\lambda,\alpha) &= \prod_{i=1}^{n}[f(X_i|(\lambda,\alpha))]^{\delta_i}[S(X_i|(\lambda,\alpha))]^{1-\delta_i} \\ &= \prod_{i=1}^{n}[\lambda\alpha X_i^{\alpha-1}e^{-\lambda X_i^{\alpha}}]^{\delta_i}[e^{-\lambda X_i^{\alpha}}]^{1-\delta_i} \\ &= \prod_{i=1}^{n}[\lambda\alpha X_i^{\alpha-1}]^{\delta_i}[e^{-\lambda X_i^{\alpha}}] \end{aligned}$$

It is important to keep in mind that originally this likelihood function $L_n(\lambda, \alpha)$ involves with corresponding parts from the censoring mechanism characterized by $g(C_i) = G'(C_I)$:

$$\prod_{i=1}^{n}[g(X_i|(\lambda,\alpha))]^{1-\delta_i}[(1-G(X_i|(\lambda,\alpha))]^{\delta_i}$$

This corresponding part is to be a product term in the likelihood function due to the independence of censoring. It is omitted due to the fact that it doesn't contain information of parameters of interest. It would be omitted if it does.

Then the log-likelihood function $l_n(\lambda, \alpha) = \log L_n(\lambda, \alpha)$ is used to construct the score equations and sample Fisher's information matrix of (λ, α) as follows:

$$\begin{aligned} s(\lambda,\alpha) &= (s_\lambda, s_\alpha)'; \\ s_\lambda &= \frac{\partial}{\partial\lambda} l_n(\lambda,\alpha); \\ s_\alpha &= \frac{\partial}{\partial\alpha} l_n(\lambda,\alpha); \end{aligned}$$

Thus the MLE of (λ, α) is calculated by solving the system of score equaltions:

$$\widetilde{0} = s(\hat{\lambda}, \hat{\alpha})$$

with a Fisher's information matrix calculated as:

$$\begin{aligned} i(\lambda,\alpha) &= \begin{bmatrix} i_{\lambda^2} & i_{\lambda,\alpha)} \\ i_{\lambda,\alpha} & i_{\alpha^2} \end{bmatrix} \\ i_{\lambda^2} &= (-1)\frac{\partial^2}{\partial\lambda^2} l_n(\lambda,\alpha; \\ i_{\lambda,\alpha} &= (-1)\frac{\partial^2}{\partial\lambda,\partial\alpha} l_n(\lambda,\alpha); \\ i_{\alpha^2} &= (-1)\frac{\partial^2}{\partial^2\alpha} l_n(\lambda,\alpha); \end{aligned}$$

and evaluated at $i(\hat{\lambda}, \hat{\alpha})$ such that the following asymptotical result holds based on Central limit Theorem in general

$$[(\hat{\lambda}, \hat{\alpha})' - (\lambda, \alpha)'] \sim N(\widetilde{0}, i^{-1}(\hat{\lambda}, \hat{\alpha})).$$

1.4 Minimum Chi-square approach.

Given right censored data $\{(X_i, \delta_i)\}_1^n$, we construct the Nelson-Aalen estimate of the cumulative hazard function $\hat{\Lambda}_{NA}(t)$ and derive a histogram based on it with the bin boundaries denoted as $\{(a_h, b_h]\}_{h=1}^H$ with $a_{h+1} = b_h$ for all $h = 1, \ldots, H$. Ideally each bin is characterized by a constant hazard function as its specific characteristics of the observed data. And the asymptotic theory assures us that: across all h's, we have independent increments

$$\sqrt{n}[\Delta_h \hat{\Lambda}_{NA}(t) - \Delta_h \Lambda(t)] \sim N(0, \int_{a_h}^{b_h} \frac{dF_u(y)}{[1 - H(y)]^2}).$$

That is, we have a Chi-square distribution with H degree of freedom as

$$n \sum_{h=1}^{H} \frac{[\Delta_h \hat{\Lambda}_{NA}(t) - \Delta_h \Lambda(t)]^2}{\int_{a_h}^{b_h} \frac{dF_u(y)}{[1-H(y)]^2}} \sim \chi^2_H$$

If T_i is Weibull distributed according to $W(\lambda, \alpha)$ with $\Lambda_W(t) = \lambda t^\alpha$, then we have the chi-square function defined as

$$\chi^2(\lambda, \alpha) = \sum_{h=1}^{H} \frac{[\Delta_h \hat{\Lambda}_{NA}(t) - (\lambda b_h^\alpha - \lambda a_h^\alpha)]^2}{\sum_{a_h \le X_{(i)} \le b_h} \frac{\delta_{(i)}}{(n-i)(n-i+1)}}$$

achieving its minimum around the true parameter values (λ_0, α_0) with

$$\chi^2(\lambda_0, \alpha_0) \sim \chi^2_H.$$

The minimum chi-square estimate of (λ_0, α_0) is computed as:

$$(\hat{\lambda}_{MC}, \hat{\alpha}_{MC})' = \arg_{(\lambda,\alpha)} \min \chi^2(\lambda, \alpha);$$

And we also have that the legitimate goodness-of-fit testing statistics

$$\chi^2(\hat{\lambda}_{MC}, \hat{\alpha}_{MC}) \sim \chi^2_{H-2}.$$

The key point of implementing minimum Chi-square approach is to perform estimation and modeling checking all at the same time. It is important to remark that the Chi-square statistics at the MLE $(\hat{\lambda}, \hat{\alpha})$ is not χ^2_{H-2} distributed.

Homework#4-3. Perform minimum Chi-square estimation and Goodness-of-fit testing on your real data set.

References

[1] Hsieh Fushing. and Roy T. Complexity of Possibly-gapped Histogram and Analysis of Histogram (ANOHT). *Royal Socity-Open Science*, 2018.

[2] Hsieh Fushing, Liu S.-Y., Hsieh Y.-C., and McCowan B. From patterned response dependency to structured covariate dependency: categorical-pattern-matching. *PLoS One*,2018.

Lectures of Machine learning in Survival analysis.

Hsieh Fushing *
Department of Statistics, University of California, Davis.

October 31, 2018

1 Lecture 0: Mimicking

In Data Science and Machine Learning, the concept of mimicking plays an essential role in inference. The skills to be able to mimic one sample of data or one data matrix would be proved critical. We will appreciate its significance in this series of lecture-notes on Survival analysis dominated by a theme of machine learning.

SIMULATION If you are given a known continuous distribution function $F(t)$ on R^1, and asked to SIMULATE n random samples from $F(t)$, then you likely to use an existing algorithm, such as in R-Studio, to generate n data points, say $\{U_i\}_{i=1}^n$, from Uniform distribution $U[0,1]$. Then you transform U_i into Y_i via the inverse function: for all $i = 1, \ldots, n$,

$$Y_i = F^{-1}(U_i).$$

This operation is called simulation, not so-called mimicking.

Mimicking patterns characterizing observed Data In this Lecture-notes, we first focus on mimicking one sample of real-valued data points without censoring, say $\{X_i\}_{i=1}^n$. Denote its ordered statistics as $\{X_{(i)}\}_{i=1}^n$. The distance measure is the Euclidean one:

$$d(X_i, X_j) = |X_i - X_j|.$$

The $n \times n$ distance matrix is constructed and denoted $D = [d(X_i, X_j)]$. Based on D and a choice of module, such as complete or Word-D2, the hierarchical clustering (HC) algorithm is applied to build a clustering tree upon $\{X_i\}_{i=1}^n$. If you choose a tree-level that corresponds to a composition of H clusters, then each of these H clusters will define a bin with its left- and right-boundaries denoted as $[a_h, b_h]$ and its height calculated as

$$\hat{f}_h = \frac{p_h}{b_h - a_h}$$

with the bin's weight $p_h = \frac{\#\{X_i | a_h \le X_i \le b_h\}}{n}$. Here we have $\sum_{h=1}^H p_h = 1$.

*Correspondence: Hsieh Fushing, University of California at Davis, CA, 95616. E-mail: fhsieh@ucdavis.edu

This histogram indeed is justified as a piecewise linear approximation to the empirical distribution function $F_n(t) = \sum_{i=1}^{N} 1_{[X_i \le t]}$, see theoretical and practical details in [1]. As a remark, it takes uneven bins sizes and potential gaps to reflect the authentic patterns of the empirical distribution $F_n(t)$.

The next two steps of mimicking after constructing a histogram, which is the [M1]-step, are given as follows:

[M2 :] Partition the unite interval $[0, 1]$ into H subintervals with respect to $\{p_h\}_{h=1}^{H}$ in a cumulative fashion, that is, using $H-1$ cut-off points as $\{p_1, p_1+P_2, \ldots, \sum_{h=1}^{H-1} p_h\}$. Then generate n random data points from $U[0, 1]$ and count how many of these simulated data points falling into each subintervals. Denote the numbers of count as $\{n_1^*, n_2^*, \ldots, n_H^*\}$.

[M3 :] For all $h = 1, \ldots, H$, generate n_h^* data points from Uniform distribution $U[a_h, b_h]$ as n_h^* mimicries from the h-th bin $[a_h, b_h]$ of the histogram.

Here is an important note for the setting when data is right censored $\{(X_i, \delta_i)\}_1^n$ with $X_i = T_I \wedge C_i$ and $T_i \perp C_i$ for all $i = 1, \ldots, n$. To mimic such a data set, given the independent censoring assumption holds, two histograms are involved: one for $\{T_i\}$ and another for $\{C_i\}$. The one for $\{T_i\}$ is constructed based on the Kaplan-Meier estimate of the survival function of T_i. Likewise the one for $\{C_i\}$ is constructed based on the Kaplan-Meier estimate of the survival function of C_i. This is the so-called duality between $\{T_i\}$ and $\{C_i\}$. When two independent samples, say $\{T_i^*\}$ and $\{C_i^*\}$, are created, then the mimicry of $\{(X_i^*, \delta_i)\}_1^n$ is generated.

Once again, the independent censoring assumption have to hold to do the above mimicking. If it is tested and failed, then the mimicking approach can be involved. We tentatively leave this issue open for now.

Mimicking based on piecewise-linear approximation to cumulative hazard function. Supposed that a piecewise-linear approximation to the cumulative hazard function $\Lambda(t) = -\log\{S(t)\}$ with $S(t) = 1 - F(t)$, is constructed as:

$$\tilde{\Lambda}(t) = \int_0^t \sum_{h=1}^{H} \tilde{\lambda}_h 1_{[a_h, b_h]}(s) ds.$$

So if $t \in [a_h, b_h]$, then we have

$$\tilde{\Lambda}(t) = \sum_{j=1}^{h-1} \tilde{\lambda}_j (b_j - a_j) + \tilde{\lambda}_h (t - a_h).$$

Then a continuous survival function is constructed as:

$$\tilde{S}(t) = e^{-\tilde{\Lambda}(t)}.$$

Hence mimicking based on $\tilde{F}(t) = 1 - \tilde{S}(t)$ can be performed as in the above Simulation setting. Once again in the right censoring setting, two mimicries: one for survival time variable of interest and one for censoring time variable, are needed to constructed one mimicry sample of right censored data.

References

[1] Hsieh Fushing. and Roy T. Complexity of Possibly-gapped Histogram and Analysis of Histogram (ANOHT). *Royal Socity-Open Science*, 2018.

[2] Hsieh Fushing, Liu S.-Y., Hsieh Y.-C., and McCowan B. From patterned response dependency to structured covariate dependency: categorical-pattern-matching. *PLoS One*,2018.

Lectures of Machine learning in Survival analysis.

Hsieh Fushing *
Department of Statistics, University of California, Davis.

November 7, 2018

1 Lecture 5: Survival Analysis at the crossroad: II. Response-to-covariate association problem.

1.1 What is the problem?

The most essential issue facing scientists is seeking directed associations from response variables to covariate features. The issue primarily centers at how to evaluate the sensitivity or variability of this relationship. If this relationship is based on a man-made structure, such as linearity, then I doubt whether the sensitivity or variability concepts are truly the most relevant concept in such a scientific endeavor?

The answer is likely negative. And the answer key is heterogeneity. While heterogeneous subpopulations almost certainly partition a large enough study population, the concept of variation needs to be multiscale in order to embrace heterogeneity, that is, the presence of homogeneous variation is different from subpopulation to subpopulation. We develop and demonstrate that machine learning algorithms can effectively accommodate such a reality of heterogeneity by extracting a spectra of pattern-based directed associations from response categories to multiple identified covariate mechanisms.

At the junction of Big Data and Artificial Intelligence, scientists wonder whether any brand newly developed computing algorithm or fundamental concept can more realistically and efficiently resolve their scientific questions than existing techniques. Such a concern is legitimate and crucial because information contents or intelligences contained in larger datasets or databases can be much more diverse and complex. Further such intelligences are not only likely going beyond human's knowledge domain, but also usually exceeds the scope of what human brain can naturally process or reach. In this paper we manifest such intelligences under the setting of survival analysis with compromised right censored data.

One chief concept attached to Big Data is heterogeneity. That is, a system, from which the data comes from, almost certain embeds with multiple distinct mechanisms. Ideally the heterogeneity is a spectra of coupled mechanism-specific states. Therefore it is critical to think that all spectra of heterogeneity render the system to have characteristically distinct manifestations.

*Correspondence: Hsieh Fushing, University of California at Davis, CA, 95616. E-mail: fhsieh@ucdavis.edu

A simple example is a subtype of a cancer or disease in medicine or a species in ecology. Thus, it is natural to note that the idea of variation has to be subtype specific, or species specific.

1.2 Models

Let the data set be denoted by $\{(X_i, Z_i, \delta_i)|X_i \in R^1_+, Z_i = (Z^1_i, ..., Z^K_i) \in R^K, \delta_i = 1 or 1 i = 1, ..., n.\}$ with δ_i being the censoring status. On the covariate side, there are K time-independent features $\{Z^1, ..., Z^K\}$, while the response variable is Y's. The primary scientific interest is to the answer to the question: How X's associates with $\mathcal{Z}$'s?

Even though the K features $\{Z^1, ..., Z^K\}$ might have been selected with care, the fact is that they are selected to cover multiple potential mechanisms, not just for one, in most scientific endeavors. Therefore we need to first figure out what are potential mechanisms involving within this system manifested via $\mathcal{Z}$? This is one perspective of heterogeneity on the covariate side only. Unfortunately this step have never been seriously taken even in Statistical modeling based analysis? For instance, the Cox's proportional hazard (PH) regression model and the Accelerated failure time (AFT) model:

1 [Cox's proportional hazard (PH) regression model:] With $X_i = T_i \wedge C_i$ and given Z_i, for all $i = 1, \ldots, n$, the hazard function $\lambda_{T_i}(t)$of T_i is:

$$\lambda_{T_i}(t) = \lambda_0(t) \exp \beta' Z_i$$

Where $\lambda_0(t)$ is unknown baseline hazard function, and $\beta = (\beta_1, \ldots, \beta_K)'$.

2 [Accelerated failure time (AFT) model:]

$$\begin{aligned} U_i &= e^{\beta' Z_i} T_i \\ \log T_i &= \beta' Z_i + \epsilon_i \\ \Lambda_{T_i}(t) &= \Lambda_{U_i}(e^{\beta' Z_i} t) \\ \lambda_{T_i}(t) &= \lambda_{U_i}(e^{\beta' Z_i} t) e^{\beta' Z_i}. \end{aligned}$$

When the covariate is indeed time-dependent, denoted as $\{Z^1(t), ..., Z^K(t)\}$, then the two model is generalized as follows:

1 [Cox's proportional hazard (PH) regression model with time-dependent covariate:] With $X_i = T_i \wedge C_i$ and given Z_i, for all $i = 1, \ldots, n$, the hazard function $\lambda_{T_i}(t)$of T_i is:

$$\lambda_{T_i}(t) = \lambda_0(t) \exp \beta' Z_i(t)$$

Where $\beta = (\beta_1, \ldots, \beta_K)'$ is usually kept to be time-independent.

2 [Accelerated failure time (AFT) model with time-dependent covariate:] let all U_i be i.i.d,.

$$\begin{aligned} U_i &= \int_0^{T_i} e^{\beta' Z_i} d\Lambda_0(t) \\ \Lambda_{T_i}(t) &= \Lambda_U(\int_0^{T_i} e^{\beta' Z_i(t)} d\Lambda_0(t)) \\ \lambda_{T_i}(t) &= \lambda_U(\int_0^{t} e^{\beta' Z_i(s)} d\Lambda(s)) e^{\beta' Z_i(t)} \lambda_0(t). \end{aligned}$$

The AFT with time-dependent covariate in the form integral equation can indeed be used to deduce all the other three models mentioned above. Its interpretation of AFT time-dependent covariate is given as: Let U_i be a random reserve, or simply the e^{ϵ_i} in the semi-parametric linear regression setup of AFT with time-independent covariate, pertaining to ith subject. The rate this subject is using up the reserve is $e^{\beta' Z_i(t)}\lambda_0(t)$. This integral equation model says that the survival time T_i is the moment the U_i is used up.

This interpretation is very logical and dynamic. Its dynamics is characterized by the fact that the history $\bar{Z}_i(t) = \{Z_i(s)|s \leq t\}$ is very relevantly involving in the cumulative hazard function of T_i:

$$\Lambda_{T_i}(t) = \Lambda_U(\int_0^{T_i} e^{\beta' Z_i(t)} d\Lambda_0(t)).$$

This involvement of covariate history also brings out the sense of acceleration because the new time scale is in an integral form w.r.t. time scale of $\Lambda_U(.)$. The effect of covariate history and sense of acceleration will be lost if $\lambda_U(.)$ is a constant, that is, U has the memory-less property. This is what underlying the Cox's proportional hazard (PH) regression model with time-dependent and time-independent covariate.

It is worth emphasizing once again that this assumption underlying Cox's proportional hazard (PH) regression model with time-dependent covariate is a rather strong assumption. At the moment of time t, the risk or hazard the subject facing depending only on $Z_i(t)$ at the moment, not on its past history, is rather unrealistic. **To our view, by accommodating history $\bar{Z}_i(t) = \{Z_i(s)|s \leq t\}$ in a reasonable model is a chief way of making relevant connections to majority of real world problems.**

On the other hand, we still need to recognize that the linearity in the rate $e^{\beta' Z_i(t)}$ is man-made, so is the integral equation. We attempt to do without such man-made structures in the developments of Machine Learning for response-to-covariate association.

1.3 Inferences: estimation and goodness-of-fit

Cox's proportional hazard regression model with time independent covariate We first discuss the profile likelihood inference approach for Cox's proportional hazard regression model with time independent covariate. With the data set $\{(X_i, Z_i, \delta_i)|X_i \in R^1_+, Z_i = (Z_i^1, ..., Z_i^K) \in R^K, \delta_i = 1 or 1 i = 1, ..., n.\}$ The likelihood contributed by (X_i, Z_i, δ_i) is computed as follows:

1 If $\delta_i = 1$, then

$$L_i(\lambda_0(.), \beta) = f_{T_i}(X_i|\lambda_0(.), \beta) = \lambda_0(X_i)e^{\beta' Z_i} e^{\{-e^{\beta' Z_i}\Lambda_0(X_i)\}};$$

2 If $\delta_i =$, then

$$L_i(\lambda_0(.), \beta) = S_{T_i}(X_i|\lambda_0(.), \beta) = e^{\{-e^{\beta' Z_i}\Lambda_0(X_i)\}};$$

The likelihood of infinite and finite dimensional parameters $(\lambda_0(.), \beta)$ is

$$\begin{aligned} L_n(\lambda_0(.), \beta) &= \prod_{i=1}^{n} [L_i(\lambda_0(.), \beta)]^{\delta_i} [L_i(\lambda_0(.), \beta)]^{1-\delta_i} \\ &= \prod_{i=1}^{n} [\lambda_0(X_i) e^{\beta' Z_i}]^{\delta_i} e^{\{-e^{\beta' Z_i} \Lambda_0(X_i)\}}. \end{aligned}$$

With the presence of infinite dimensional parameter $\lambda_0(.)$, the optimization upon the likelihood function $L_n(\lambda_0(.), \beta)$ is not immediately doable. One solution to this issue is to restrict candidates of $\lambda_0(.)$ in a finite dimensional space:

$$\{\tilde{\lambda}(.) | \tilde{\lambda}(X_{(i)}^{(u)}) = \lambda_{u(i)}, i = 1, \ldots, n_u\}.$$

That is, we only consider candidates of $\lambda_0(.)$, which have discrete weights only at uncensored time points. Upon this restricted space for $\lambda(.)$, we have the following:

$$\begin{aligned} \tilde{\Lambda}(t) &= \sum_{X_{(i)}^{(u)} \le t} \lambda_{u(i)}, \\ L_n(\tilde{\lambda}(.), \beta) &= \prod_{i=1}^{n_u} \lambda_{u(i)} e^{\beta' Z_{(i)}^{(u)}} e^{\{\sum_{j=1}^{n} \tilde{\Lambda}(X_{(j)}) e^{-\beta Z_{(j)}}\}}, \\ \sum_{j=1}^{n} \tilde{\Lambda}(X_{(j)}) e^{-\beta Z_{(j)}} &= \sum_{j=1}^{n} [\sum_{X_{(i)}^{(u)} \le X_{(j)}} \lambda_{u(i)}] e^{-\beta Z_{(j)}} \\ &= \sum_{i=1}^{n_u} \lambda_{u(i)} [\sum_{X_{(j)} \ge X_{(i)}^{(u)}} e^{-\beta Z_{(j)}}] \end{aligned}$$

We then have the log-likelihood function as:

$$l_n(\tilde{\lambda}(.), \beta) = \sum_{i=1}^{n_u} (\log \lambda_{u(i)} - \beta' Z_{(i)}^{(u)}) - \sum_{i=1}^{n_u} \lambda_{u(i)} [\sum_{X_{(j)} \ge X_{(i)}^{(u)}} e^{-\beta Z_{(j)}}].$$

and the score equations for $\{\lambda_{u(i)}\}$ are calculated as:

$$0 = \frac{\partial}{\partial \lambda_{u(i)}} l_n(\tilde{\lambda}(.), \beta) = \frac{1}{\lambda_{u(i)}} - \sum_{X_{(j)} \ge X_{(i)}^{(u)}} e^{-\beta Z_{(j)}},$$

So that the profiled estimates of $\{\lambda_{u(i)}\}$ are calculated as: given any β value, for all $i = 1, \ldots, n_u$,

$$\hat{\lambda}_{u(i)} = \frac{1}{\sum_{X_{(j)} \ge X_{(i)}^{(u)}} e^{-\beta Z_{(j)}}}.$$

The profiled likelihood of β is calculated as:

$$L_n(\hat{\tilde{\lambda}}(.;\beta),\beta) = \prod_{i=1}^{n_u} \frac{e^{\beta' Z_{(i)}^{(u)}}}{\sum_{X_{(j)} \geq X_{(i)}^{(u)}} e^{-\beta Z_{(j)}}}$$

The summation range in each denominator is called "risk set". This profiled likelihood of β is termed by D.R. Cox (1972) "the partial likelihood" of *beta*. It should be noted that this profiled likelihood is a product of "potential ratios" that is commonly used in Statistical Mechanics of Physics. This turns out to be an effective way of getting around with infinite dimensional parameter.

The MLE of β based on its partial likelihood is found by solving the score equations:

$$\frac{\partial}{\partial \beta} l_n(\hat{\tilde{\lambda}}(.;\beta),\beta) = 0,$$

and its asymptotical distribution is derived and expressed as follows:

$$\begin{aligned} \hat{\beta} &\sim N(\beta, i_\beta^{-1}) \\ i_\beta &= (-1)\frac{\partial^2}{\partial \beta^2} l_n(\beta|\hat{\tilde{\lambda}}(.;\beta)) \end{aligned}$$

Also the (generalized) MLE estimate of $\Lambda(t)$ is calculated:

$$\hat{\tilde{\Lambda}}_0(t) = \sum_{X_{(i)}^{(u)} \leq t} \hat{\lambda}_{u(i)}(\hat{\beta})$$

Further since the fact that

$$Pr[\Lambda_0(T_i) > t] = Pr[T_i > \Lambda_0^{-1}(t)] = e^{-e^{\beta' Z_i}\Lambda(\Lambda_0^{-1}(t))} = e^{-e^{\beta' Z_i} t},$$

So that the transformed random variable

$$e^{\beta' Z_i}\Lambda_0(T_i) \sim EXP(1).$$

Therefore, for the purpose of Goodness-of-fit, we can make use of the transformed data set $\{(\hat{U}_i, \delta_i)\}_{i=1}^n$: for all $i = 1, \ldots, n$,

$$\hat{U}_i = e^{\hat{\beta}' Z_i}\hat{\tilde{\Lambda}}_0(X_i)$$

to build a Chi-squared testing statistics based on its Nelson-Aalen estimate and its asymptotical convergence to a Gaussian process with independent increment property.

Implied interpretations of Partial Likelihood The popular interpretation of partial likelihood offered in Cox (1972) is based on conditional probability: among the surviving subjects up to just before the next uncensored event time, say $X_{(i)}^-$, the probability of the event happened on i-th subject is calculated as:

$$\frac{e^{\beta' Z_{(i)}^{(u)}}}{\sum_{X_{(j)} \geq X_{(i)}^{(u)}} e^{-\beta Z_{(j)}}}.$$

The "ideal MLE" is the β value such that each every conditional probability achieves its possible "maximum " value from the smallest to largest uncensored event time points. Here the conditional probability as mentioned before is simply a ratio of individual potential $e^{\beta' Z_{(i)}^{(u)}}$ over total sum of potentials $\sum_{X_{(j)} \geq X_{(i)}^{(u)}} e^{-\beta' Z_{(j)}}$ belonging to the risk set.

It is crucial to bear in mind that $e^{\beta' Z_{(i)}^{(u)}}$ is not the only way of defining potential. There are many ways to assign reasonable potentials that are not necessarily relying on linearity on covariate vector values, that is, $e^{\beta' Z_i}$ is just one of the simplest versions. On the other hand, the linearity might go against the most intuitive Euclidean closeness, that is, Z_i and Z_j being far-away in terms of Euclidean distance could be very close, or even identical via linearity of βZ_i. That is, Euclidean distance measure could also bring out one conceptual potential among all $\{Z_i\}$. Therefore it is reasonable to build a clustering tree upon the row-axis of $n \times K$ data matrix $\mathcal{Z} = [Z_i^k]$.

Though this Euclidean similarity or distance based potential is seemingly very intuitive and reasonable, it is attached with one drawback that all K-features $\{Z_i^1, ..., Z_i^K\}$ in Z_i play equally important roles. This requirement of equal importance among all involving feature is not likely to be true. We need some more developments to lessen possible impacts from this drawback.

Homework $\#5-1$. Perform the Cox's proportional hazard regression model on your data set. If your data set is of the type of K-sample problem, then you need to define the covariate Z_i to transform the K-sample problem into a hazard regression one. And perform the Goodness-of-fit testing as well.

Homework $\#5-2$. Construct a histogram based on $\{(X_i, \delta_i)\}$ (ignoring the Z_i) and color-code all bins. Further construct a histogram based on $\{e^{\hat{\beta}' Z_i}\}$, and display the color contents on each of its bins. Build a clustering tree upon the row-axis of $n \times K$ data matrix $\mathcal{Z} = [Z_i^k]$, and then choose a tree-level to identify a composition of clusters and color-code each cluster in the clustering composition.

Lectures of Machine learning in Survival analysis.

Hsieh Fushing *
Department of Statistics, University of California, Davis.

November 27, 2018

1 Lecture 6: Survival Analysis at the crossroad: II. Response-to-covariate association problem (Cont'd).

1.1 Cox' proportional hazard regression model with time-dependent covariate.

Recall Cox's proportional hazard (PH) regression model with time-dependent covariate: With $X_i = T_i \wedge C_i$ and given the covariate history $\bar{Z}_i(t) = \{Z_i(s) = (Z^1(s), ..., Z^K(s)) | 0 < s \leq t\}$, for all $i = 1, \ldots, n$, the hazard function $\lambda_{T_i}(t)$ of T_i is:

$$\lambda_{T_i}(t) = \lambda_0(t) \exp \beta' Z_i(t)$$

where $\beta = (\beta_1, \ldots, \beta_K)'$ is usually kept to be time-independent. It is important to emphasize again that, at time t, this model assumption restricts the amount of risk (via hazard function) facing a subject depends on only $Z_i(t)$, not the entire history $\bar{Z}_i(t)$. This is certainly very unrealistic. For inference purpose, the partial likelihood approach is commonly applied here: the likelihood of β is calculated as:

$$L_n(\beta) = \prod_{i=1}^{n_u} \frac{e^{\beta' Z_{(i)}^{(u)}(X_{(i)}^{(u)})}}{\sum_{X_{(j)} \geq X_{(i)}^{(u)}} e^{\beta Z_{(j)}(X_{(i)}^{(u)})}}.$$

Immediately it is natural to ask:

Homework $\#6 - 1$. Is this partial likelihood a version profiled likelihood as in the time-independent covariate setting? This area of Survival Analysis is also called "Event-history analysis", in particularly, when Martingale theory is employed [1] [2].

*Correspondence: Hsieh Fushing, University of California at Davis, CA, 95616. E-mail: fhsieh@ucdavis.edu

1.2 AFT model: regression with censored data

After considering the Cox' proportional hazard regression model, we now turn to AFT model with time independent covariates. AFT model model on the data set $\{(X_i, Z_i, \delta_i)|X_i \in R^1_+, Z_i = (Z_i^1, ..., Z_i^K) \in R^K, \delta_i = 1 or 1 i = 1, ..., n.\}$ with δ_i being the censoring status is described as: for all $i = 1, \ldots, n$,

$$\log T_i = \beta' Z_i + \epsilon_i$$

with $\epsilon_i = \log U_i$. Since U_i is completely unknown distributed, this model is in the form of semiparametric linear regression setting. It is further complicated by a right censoring mechanism. Therefore the likelihood approach is not directly applicable because the presence of acceleration component $e^{\beta' Z_i}$ working on time scale t via:

$$\Lambda_{T_i}(t) = \Lambda_{U_i}(e^{\beta' Z_i} t).$$

This makes the significant difference between Cox's model and AFT. That is, the profiled likelihood approach will not work in AFT setting. Different approaches have to be devised.

One most common approach used in semiparametric regression setting is to make use of the property of Z_i and ϵ_i being uncorrelated:

$$\widetilde{Cov}(Z_i, \epsilon_i) = \tilde{0}$$

with $\tilde{0} \in R^K$.

Let's denote $Y_i = \log X_i$ for all i's. Then we have that, if there exist no censored data points,

$$\begin{aligned} \sum_{i=1}^n (Z_i - \bar{Z}_n) Y_i &= \sum_{i=1}^n (Z_i - \bar{Z}_n) Z_i^T \beta + \sum_{i=1}^n (Z_i - \bar{Z}_n) \epsilon_i \\ &\approx n \hat{\Sigma}_Z \beta + \tilde{0} \end{aligned}$$

with $\hat{\Sigma}_Z$ denoting the sample covariance computed from $\{Z_i\}_{i=1}^n$. Therefore we can have a semiparametric estimate of β calculated as:

$$\hat{\beta} = [n \hat{\Sigma}_Z]^{-1} \sum_{i=1}^n (Z_i - \bar{Z}_n) Y_i.$$

To make use of this approach, we need to impute X_i for all censored data point, that is, we want to estimate $\log T_i$ when $\delta_i = 0$ in the following fashion as: if β is known, then

$$Y_i^* = \hat{E}[\log T_i | T_i > X_i] = Z_i^T \beta + \hat{E}[\epsilon_i | \epsilon_i > \log X_i - Z_i^T \beta].$$

This brutal force approach needs a starting estimate $\hat{\beta}_0$, and then we estimate the residuals as: for all is,

$$V_i = \log X_i - Z_i^T \hat{\beta}_0.$$

With estimated residuals $\{(V_i, \delta_i)\}_{i=1}^n$, we calculate the corresponding survival function $\hat{S}_\epsilon(t)$ for all $t \in R^+$.(For positivity, we might need to make a translation). And then we compute

$$\hat{E}[\epsilon_i | \epsilon_i > V_i] = \frac{\sum_{V_{(k)} \geq V_i} W_{(k)}(\hat{\beta}_0) V_{(k)}}{\hat{S}_\epsilon(V_i)}.$$

where $\{W_{(k)}(\hat{\beta}_0)\}$ is the collection of weights of jumps in $\hat{S}_\epsilon(t)$.

Then we need to update the imputed Y_i^* for all censored data points in the semiparametric estimation of β. It becames an iterative procedure, so-called Buckley-James procedure proposed by Buckle and James (1979, Biometrika). Again this is a brutal force approach, it became even more so with their proposal of a variance estimate as:

$$\begin{aligned} \hat{Var}(\hat{\beta}) &= [n\hat{\Sigma}_Z]^{-1}\hat{\sigma}_u^2 \\ \hat{\sigma}_u^2 &= \frac{1}{n_u - 2}\sum_{i=1}^{n_u}(T_i - \bar{T}_u - \hat{\beta}(Z_i - \bar{Z}_u))^2 \end{aligned}$$

with $\sum_{i=1}^{n_u}$ summing over uncensored data points and $\bar{T}_u$ computed as the average of uncensored survival times.

Homework $\#6-2$. Compare results based on Cox's hazard regression model and accelerated failure time (AFT) model on your data set.

1.3 Reflections

A more subtle and elegant estimation, called Empirical Process Approach (EPA), was proposed and studied in Fushing (1997, Annals of Statistics). The basis is the counting process' martingale central limit convergence theory. In the paper the point and interval estimations and goodness-of-fit testing are simultaneously dealt with and resolved.

The inferences for integral equation model with time-dependent covariate discussed in Lecture-notes-5 are completely resolved with semiparametric efficiency are reported in Fushing (2012, Annals of Institute of Statistical Mathematics). A system of counting processes is constructed, and the martingale central limit theory again played the essential role in building the asymptotical results.

After 20 years, it might be time to ask a necessary question on reflecting modeling as one whole issue: Is such a modeling indeed conceptually proper? More specifically, why and how the different response categories potentially coupled with rather distinct covariate clusters are governed by a single homogeneous additive error mechanism?

References

[1] T. R. Fleming and D. P. Harrington Counting Process & Survival Analysis. *Wiley-Interscience, New York*, 1991.

[2] P.K. Andersen, O. Borgan, R. D. Gill and N. Keiding Statistical Models Based on Counting Process. *Springer-Verlag, New York*, 1993.

[3] Hsieh Fushing. and Roy T. Complexity of Possibly-gapped Histogram and Analysis of Histogram (ANOHT). *Royal Socity-Open Science*, 2018.

[4] Hsieh Fushing, Liu S.-Y., Hsieh Y.-C., and McCowan B. From patterned response dependency to structured covariate dependency: categorical-pattern-matching. *PLoS One*,2018.