Nonparametric analysis of recurrent events with incomplete observation gaps

Jinheum Kim
University of Suwon
Korea
Recurrent event data

- Recurrent events arise frequently in clinical trials in which patients are followed longitudinally and response of interest is transient.

- Examples
  - Transient ischemic attacks in patients with cerebrovascular disease
  - Seizures in epileptic patients
  - Tumor recurrences in cancer patients
Motivating example

- Young Traffic Offenders Program (YTOP) data

- What is YTOP?
  - Conducted at the University of Missouri-Columbia Health Sciences Center and other medical centers since late 1987
  - One-day educational intervention for people under 24 convicted of speeding by 20+ mph over the posted speed limit
To evaluate the effect of the program on reducing the rates of speeding violation of young drivers
Data were collected on 192 young drivers about their speeding violation information since they obtained their driving license.

29 subjects (13: YTOP, 16: non-YTOP) received suspensions.

These subjects dropped out of the study when the suspension began, and then came back to the study after completing suspension.

These intermittent dropouts are called **observation gaps**.

The observation gap is determined by the starting and terminating time of the suspension.

**BUT**, the observation gaps of YTOP data are **incomplete** due to missing terminating times.
Previous works & their limitations

- Farmer et al. (2000, Brain Injury): two-sample rank test
  - Ignore the detailed information about the conviction process
  - Can assess only the long-term impact

- Sun et al. (2001, JASA): nonparametric & semi-parametric methods
  - Treat YTOP data as recurrent event data
  - Evaluate both short- and long-term effects applying the conviction process information
  - Ignore the observation gaps
Previous works & their limitations

- Zhao & Sun (2006, CSDA)
  - Similar to Sun et al. (2001, JASA), but consider the observation gaps
  - Suspension periods: NOT deterministic, BUT varying and possibly unknown

  - Treat incomplete observation gaps as terminating times being interval-censored
  - Only use the first suspension and ignore all the other suspensions following the first one.
Objective of this talk

- Use the multivariate-interval censored data to estimate the distribution of the terminating time

- Propose a non-parametric test to compare the conviction rates of two groups

- Evaluate the proposed test via simulation & reanalyze the YTOP data
Illustrative data: notation

- ●: conviction time
- X: starting time of the suspension
- ?: terminating times of the suspension (unknown)
- ■: end of study time
Diagram
Want to estimate the distribution of the terminating time using the interval-censored data to incorporate the observation gaps
Two types of IC data

- **Univariate IC data**
  - cases II, III, & V having one observation gap
  - eg: $(5,9]$, $(7, \infty)$, $(2,10]$

- **Multivariate IC data**
  - case IV having two observation gaps
  - eg: $(2,5]$. BUT, …
How can we define the origin of the second terminating time?
Constructing IC data

- Set the origin to be the first conviction time after the previous suspension
  - eg: For the second suspension of case IV,

  \[(7 - 5 = 2, 9 - 5 = 4]\]

- Hence, estimate the distribution of the terminating time using

  \[(5, 9], (7, \infty), (2, 10], (2, 5] + (2, 4]\]
Hybrid algorithm (Wellner & Zhan, 1997, JASA)

Use R package interval
Notation & assumption

- \( N_i(t) \) : the number of occurrences of the event up to time \( t \), \( i = 1, \ldots, n \)

- \( Y_i(t) \) : function indicating if subject is under observation at time \( t \)

- \( \Lambda_i(t) = E\{N_i(t)\} \) cumulative mean (frequency) function (CMF)

- \( N_i \& Y_i \) independent
Data

- $D = \{(t_{ik}, \tilde{Y}_i, h_{ij}), i = 1, \ldots, n; k = 1, \ldots, n_i; j = 1, \ldots, m_i\}$

- $t_{ik}$: the $k$th recurrent event time of subject $i$

- $\tilde{Y}_i$: 0-1 indicator whether or not subject $i$ has experienced observation gap

- $h_{ij}$: starting time of the $j$th observation gap of subject $i$
Unobservable terminating times

- BUT, the terminating time of each observation gaps is NOT available, that is, interval-censored

- So, need to estimate the distribution of the terminating time of the observation gap
IC data

- \( I = \{(L_{ij}, R_{ij}], i = 1, \ldots, n; j = 1, \ldots, m_i\} \)
- \( L_{ij} = h_{ij} - t_{ij0}, R_{ij} = t_{ij1} - t_{ij0} \)

- \( t_{ij0} \) : minimum of greater than \( h_{i,j-1} \)
- \( t_{ij1} \) : minimum of greater than \( h_{ij} \)
- \( R_{ij} \) can be right-censored
Terminating time distribution

- $S_{ij}$: the terminating time for the $i$th observation gap of subject $i$
- $a_l$: the mid-point of the $l$th equivalence class
- $f_l = P(S_{ij} = a_l)$: mass of $S_{ij}$ at $a_l$
Redefine a risk set

Replace $Y_i(t)$ by the estimated expected risk set defined as

$$Y_i^*(t) = 1 - \tilde{Y}_i \sum_{j=1}^{m_i} \sum_{l=1}^{q} \frac{J_{ij}(t)I(t-t_{ij0} \leq a_l \leq R_{ij}) \hat{f}_l}{\sum_{j=1}^{m_i} \sum_{r=1}^{q} J_{ij}(t)I(L_{ij} < a_r \leq R_{ij}) \hat{f}_r}$$
Equivalence classes

Case II: \( (L \quad R) \)
Case III: \( (L) \)
Case IV: \( (L \quad R) \)
(\(L \quad R) \)
Case V: \( (L) \quad R \)
Combine \( \quad L \quad R \quad L \quad L \quad R \quad R \)

Equivalence classes: \( (2,4\], \quad (5,5\], \quad (7,9\] 
Mid-points: 3, 5, 8

Assume \( P(S = 3) = f_1, P(S = 5) = f_2, P(S = 9) = f_3 \), with \( f_1 + f_2 + f_3 = 1 \)
Unadjusted (naive) risk set

Case I

Case II

Case III

Case VI

Case V
Modified (proposed) risk set

\[ p = \frac{f_1}{f_1 + f_2}, \quad q = \frac{f_1}{f_1 + f_2 + f_3}, \quad r = \frac{f_1 + f_2}{f_1 + f_2 + f_3} \]
Estimated CMF

- Assuming that all subjects have the common CMF,

\[ \Lambda_1(t) = \Lambda_2(t) = \cdots = \Lambda_n(t) = \Lambda(t), \]

\[ \hat{\Lambda}(t) = \int_0^t \frac{dN^*(s)}{Y^*(s)} \]

- \( dN^*_t(t) = \sum_{i=1}^n Y^*_i(t) dN^*_i(t) \): total number of events observed at time \( t \)

- \( Y^*_t(t) = \sum_{i=1}^n Y^*_i(t) \): total number of subjects at risk at time \( t \)
Representation of estimated CMF

- $t_1 < t_2 < \cdots < t_d$ distinct event times across all individuals

$$\hat{\Lambda}(t) = \sum_{j:t_j \leq t} \frac{dN^*_j(t_j)}{Y^*_j(t_j)}$$

- Similar to the Nelson-Aalen estimator from survival analysis
- Cook & Lawless (2007)
Asymptotic properties

- Following Lin et al. (2001, JRSSB),
  - Consistent
  - Asymptotically normal with consistently estimated variance,

\[
\hat{\text{var}}\{\hat{\Lambda}(t)\} = \sum_{i=1}^{n} \left\{ \sum_{j: t_j \leq t} \frac{Y_i^*(t_j)}{Y_i^*(t_j)} \left[ dN_i(t_j) - \frac{dN^*_i(t_j)}{Y_i^*(t_j)} \right] \right\}^2
\]

- Approximate \( 1 - \alpha \) pointwise CI:
\[
\exp(\log \hat{\Lambda}(t) \pm z_{\alpha/2} \sqrt{\hat{\text{var}}\{\log \hat{\Lambda}(t)\}})
\]
Comparison of two groups

- $\Lambda_1(t), \Lambda_2(t)$: the CMFs for group 1 and 2
- $n = n_1 + n_2$: the number of subjects in the $g$th group with $n_1, n_2$
- $\hat{\Lambda}_1(t), \hat{\Lambda}_2(t)$: the estimated CMF for the $g$th group
Two-sample test statistic

- **Hypothesis:** \( H_0 : \Lambda_1(t) = \Lambda_2(t), \forall t > 0 \)

- **Test statistic:**
  
  \[
  T_w = \int_0^\tau w(s) \{d\hat{\Lambda}_1(s) - d\hat{\Lambda}_2(s)\}
  \]

  - \( \tau (> 0) \): maximum follow-up time across two groups
  - \( w(t) \): non-negative predictable function
Asymptotic distribution

Following Cook et al. (1996, BCS) & Lin et al. (2001, JRSSB),

Asymptotically normal with consistently estimated variance,

\[ \hat{\text{var}}(T_w) = \sum_{g=1}^{2} \sum_{k=1}^{n_g} \left[ \int_0^\tau w(s) \frac{Y_{gk}^*(s)}{Y_{g.}^*(s)} \{dN_{gk}(s) - d\hat{\Lambda}_g(s)\} \right]^2 \]

Hence, \( T_w^2 / \hat{\text{var}}(T_w) \sim \chi^2(1) \) asymptotically
Simulation: data generation

- Generate the first event time from an exponential distribution with a hazard rate of $\lambda_i$.

- Generate $\delta_{i1} \sim B(1, p)$
  - $p$: probability of having an observation gap.
Simulation: data generation

- If $\delta_{i1} = 1$, generate the starting time of the first observation gap from $U(t_{i1}, t_{i1} + 005)$.

- If $\delta_{i1} = 1$, generate the duration time from an exponential distribution with a hazard rate of 2 resulting in the terminating time $h_{i1} + v_{i1}$.

- So, we get $t_{i1}$ and $h_{i1}$ (if $\delta_{i1} = 1$).
Simulation: the first cycle of data generation

w/ observation gap

w/o observation gap

Joint 2011 Taipei Symposium

2011-12-17
Two-sample tests: design params

- Event rate of each group
  - $\lambda_1 = 1.5$
  - $\lambda_2 = 1.5, 1.4, 1.3, 1.2, 1.1$
- Weight: $w(t) = 1$
- $p = 0.05, 0.1, 0.2, 0.4$
- Fixed censoring at 5
- Sample size: $n = 50, 100$
- Replication: 1,000
Two-sample tests: results (n=100)

<table>
<thead>
<tr>
<th>$p$</th>
<th>$\lambda_2$</th>
<th>Tests</th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Proposed</td>
<td>Naive</td>
<td>Sun et al.</td>
<td>Farmer et al.</td>
<td></td>
</tr>
<tr>
<td>0.05</td>
<td>1.5</td>
<td>0.047</td>
<td>0.055</td>
<td>0.047</td>
<td>0.047</td>
<td></td>
</tr>
<tr>
<td></td>
<td>1.4</td>
<td>0.271</td>
<td>0.273</td>
<td>0.263</td>
<td>0.190</td>
<td></td>
</tr>
<tr>
<td></td>
<td>1.3</td>
<td>0.755</td>
<td>0.740</td>
<td>0.740</td>
<td>0.593</td>
<td></td>
</tr>
<tr>
<td></td>
<td>1.2</td>
<td>0.978</td>
<td>0.972</td>
<td>0.979</td>
<td>0.918</td>
<td></td>
</tr>
<tr>
<td>0.1</td>
<td>1.5</td>
<td>0.060</td>
<td>0.055</td>
<td>0.056</td>
<td>0.039</td>
<td></td>
</tr>
<tr>
<td></td>
<td>1.4</td>
<td>0.246</td>
<td>0.231</td>
<td>0.248</td>
<td>0.182</td>
<td></td>
</tr>
<tr>
<td></td>
<td>1.3</td>
<td>0.704</td>
<td>0.668</td>
<td>0.677</td>
<td>0.538</td>
<td></td>
</tr>
<tr>
<td></td>
<td>1.2</td>
<td>0.981</td>
<td>0.966</td>
<td>0.980</td>
<td>0.907</td>
<td></td>
</tr>
<tr>
<td>0.4</td>
<td>1.5</td>
<td>0.062</td>
<td>0.049</td>
<td>0.068</td>
<td>0.054</td>
<td></td>
</tr>
<tr>
<td></td>
<td>1.4</td>
<td>0.190</td>
<td>0.151</td>
<td>0.179</td>
<td>0.138</td>
<td></td>
</tr>
<tr>
<td></td>
<td>1.3</td>
<td>0.633</td>
<td>0.440</td>
<td>0.576</td>
<td>0.423</td>
<td></td>
</tr>
<tr>
<td></td>
<td>1.2</td>
<td>0.941</td>
<td>0.793</td>
<td>0.914</td>
<td>0.772</td>
<td></td>
</tr>
</tbody>
</table>
YTOP data: estimated CMFs

(A) Proposed: YTOP

(B) Proposed: Gender

Joint 2011 Taipei Symposium
## YTOP data: p-values

<table>
<thead>
<tr>
<th>Tests</th>
<th>Covariate</th>
<th>YTOP</th>
<th>Gender</th>
</tr>
</thead>
<tbody>
<tr>
<td>Proposed</td>
<td>YTOP</td>
<td>0.836</td>
<td>0.135</td>
</tr>
<tr>
<td></td>
<td>Gender</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Naive</td>
<td>YTOP</td>
<td>0.735</td>
<td>0.161</td>
</tr>
<tr>
<td></td>
<td>Gender</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Sun et al.</td>
<td>YTOP</td>
<td>0.456</td>
<td>0.008</td>
</tr>
<tr>
<td></td>
<td>Gender</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Farmer et al.</td>
<td>YTOP</td>
<td>0.113</td>
<td>0.533</td>
</tr>
<tr>
<td></td>
<td>Gender</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Propose the expected risk set to incorporate the observation gaps

Our two-sample test shows to satisfy the nominal level & has the higher power than the other tests

No significant difference in traffic conviction rates between YTOP and non-YTOP participants & between male and female
A regression model to include a time-dependent covariate $Z(t) = I(T \leq t)$, where $T$ denotes the time at which a subject undergoes the traffic education program

$$\Lambda_i(t) = E\{N_i(t) \mid z_i(t)\} = \lambda_0(t) \exp\{\beta'z_i(t)\}$$

A case to allow a termination event such as death
Thank you!