Survival Analysis in the 21st Century: Rewriting Event History with Latent Variables

Katherine E. Masyn, Ph.D.
Graduate School of Education
Harvard University

Texas Tech University
December 6, 2013

Overview

- Survival Basics
- Discrete-time Survival Analysis (DTSA)
- Vintage DTSA
- Modern DTSA
- DTSA as SEM
- DTSA as LTA
- DTSA as LCA

Survival Basics

Time-to-event data

Event history: A record of whether and when an event occurred for each individual in a sample during a finite observation period, e.g., time of death, grade of school dropout, age of first alcohol use in school-aged children, etc.
Example

- One of the oft-cited factors that influence the quality and equality of public education in the U.S. is the lack of well-trained, high-quality teachers.
- Originally, this problem was classified at a supply-side issue, i.e., more people need to be recruited and effectively inducted into the teaching profession.
- Then this problem was classified as a retention issue.
- To further understand teacher retention/attrition and its correlates, we need a way to model both whether and when individuals exit the teaching profession.

Key Information

- What is the target event, i.e., for what is the individual at-risk?
- What defines risk onset, i.e., t = 0:00:00?
- What is the metric for measuring time from risk onset to event occurrence?
- Under what circumstances is the event time of an individual unknown or not observed, i.e., missing?

Target event

- An event for which each person in the population is theoretically at-risk, i.e., has a non-zero probability of experiencing the event at some point in time.
  - At what age
    - Did you first drink alcohol without parent permission?
    - Were you suspended from school for first time?
    - Were you first arrested by the police?
  - The risk for the event as well as the actual timing of the event varies across individuals.

Risk onset

- Start the clock when everybody is initially at-risk to experience the event
  - At what age
    - Did you first drink alcohol without parent permission? → Time 0 = Birth
    - Were you suspended from school for first time? → Time 0 = age of first school enrollment.
    - Were you first arrested by the police? → Time 0 = ?
  - Start the clock at the occurrence of a precipitating event
<table>
<thead>
<tr>
<th>Time metric</th>
</tr>
</thead>
</table>
| **Continuous**  
The “exact” time of an event for each subject is known, e.g., time of death |
| **Discrete**  
1) The timing of an event is continuous but is only recorded for an interval of time, e.g., grade of school dropout.  
2) The timing of an event is itself discrete, e.g., grade retention. |

<table>
<thead>
<tr>
<th>Missing data</th>
</tr>
</thead>
</table>
| Missing data is endemic in longitudinal studies; survival studies are no exception.  
Various mechanisms for missing data in the survival context are referred to under the unifying term, censoring, indicating that the event times for some subjects are unknown to the researcher.  
Censoring is usually assumed to be noninformative which means that the distribution of censoring times is independent of event times, conditional on the set of observed covariates. (Think: Missing-at-random) |

<table>
<thead>
<tr>
<th>Example (cont’d)</th>
</tr>
</thead>
</table>
| The most typical survival data is right-censored and this type of missingness is the easiest to deal with in the data analysis.  
Right censoring occurs when a subject in the sample has not experienced the event of interest at the end of the observation period. It is assumed that the event eventually occurs sometime after the end of the study. |
| The target event is first departure from full-time classroom teaching.  
Risk for all subjects begins on the morning of the first day of the first year of teaching.  
Events are assumed to be discretely occurring between school years.  
Once an individual has left teaching for the first time, he/she is no longer at-risk.  
Event time is unknown if individual is lost to follow-up or the time is greater than 5 years. |
Discrete-time Survival Analysis (DTSA)

Survival probability

Let $T$ be the time interval of the event where $T \in \{1,2,\ldots,J\}$

$P_S(j)$, called the survival probability, is defined as the probability of “surviving” beyond time interval $j$, i.e., the probability that the event occurs after interval $j$:

$$P_S(j) = P(T > j).$$

Hazard probability

$P_h(j)$, called the hazard probability, is defined as the probability of the event occurring in the time interval $j$, provided it has not occurred prior to $j$:

$$P_h(j) = P(T = j \mid T \geq j).$$

Essentially, $P_h(j)$ is the probability of the event occurring in time interval $j$ among those at-risk in $j$. 

Survival for full-time teaching

![Graph showing survival probability over time intervals for full-time teaching.](Image)
Use of Hazard Probabilities

- Identifies especially risky time periods, i.e., times when the event is particularly likely to occur.
- Characterizes the shape of the hazard function – determining whether risk increases, decreases, or remains constant over time.
- Most survival models are specified in terms of the hazard probabilities.

Hazard and Survival Probabilities

The relationship between $P_{S}(j)$ and $P_{h}(j)$ is given by

$$P_{S}(j) = P(T > j) = P(T > j \mid T \geq j) \cdot P(T > j - 1 \mid T \geq j - 1) \cdot \cdots P(T > 1 \mid T \geq 1)$$

$$= \prod_{k=1}^{j} (1 - P_{h}(k))$$

- When the hazard is high, the survival function drops more rapidly.
- When the hazard is low, the survival function drops slowly.
- When hazard is zero, the survival function is constant.
- Both $P_{h}$ and $P_{S}$ give us important information: $P_{S}$ tells who is at risk at a given time and $P_{h}$ tells what their risk is.
Vintage DTSA

Modern DTSA

Proportional Hazard Odds Model

\[ \log \left( \frac{P_h(j)}{1 - P_h(j)} \right) = \alpha_j + X\beta \]

Estimated using MLE within a traditional logistic regression or fixed-effects multilevel logistic regression framework.

\[ f= \text{continuous latent variable}; \ c=\text{categorical latent variable} \]
\[ y= \text{continuous observed variable}; \ u=\text{categorical observed variable} \]
\[ T= \text{continuous survival time}; \ x=\text{observed covariate} \]
Advantages of Unifying Framework

- ML estimation
- Allows for different measurement models for predictors
- Allows for multilevel data structure
- Facilitates the modeling of joint processes
  - Accounts for shared variance between processes—observed and unobserved
  - Allows for moderation or mediation of observed covariates by process-level (latent) variables
  - More accurately reflects the complexity and interconnectedness of longitudinal processes

- Easily accommodates extensions:
  - Time-varying covariates
  - Non-proportional hazard, i.e., time-varying covariates effects
  - Random effects, i.e., individually-varying covariate effects
  - Recurring events
  - Competing risks

DTSA as SEM
DTSA as SEM Extensions

- Multivariate DTSA
  - Recurring Events
  - Multiple Spell
  - Competing risk
- Event history outcomes in mediation model
- Multilevel DTSA
- Semi-parametric frailty models
- Latent class predictors of survival

<table>
<thead>
<tr>
<th>$i$</th>
<th>$e_1$</th>
<th>$e_2$</th>
<th>$e_3$</th>
<th>$e_4$</th>
<th>$e_5$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td></td>
<td>.</td>
</tr>
<tr>
<td>2</td>
<td>0</td>
<td>0</td>
<td></td>
<td>.</td>
<td>.</td>
</tr>
<tr>
<td>3</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

$\hat{P}_h(j) = \hat{P}(E_j = 1)$

Multivariate DTSA

- Commonly there is only one event state but more states are possible.
- Event type
  - Single, non-recurring
  - Recurrent events (multiple spells)
    - Same outcome that may occur more than once
  - Competing risks
    - More than one possible outcome
  - Parallel process
    - Multiple event processes occurring at the same time with independent event-specific risk status but shared variance

Recurring Event: Professional role changes for educators, e.g., classroom teacher, administrator, researcher, etc.

Research question: What is the relationship between time spent as a classroom teacher and time spent in subsequent roles in or outside of Education?
Competing Risks: Professional roles following departure from full-time classroom teaching, e.g., administrator, researcher, etc.

Research question: How does administrative support influence the hazard of leaving teaching for another role in Education compared to the hazard of leaving teaching for a professional role outside Education?

Mediation w/ Event History Outcome

Research question: How does first grade aggression mediate the effect of gender on time to first school removal?
Multilevel DTSA

- Event: First school removal (suspension or expulsion)
- Research question: How does individual-level aggressive-disruptive behavior and classroom-level aggressive-disruptive behavior influence the hazard of school removal?

Unobserved heterogeneity

Often referred to as frailty in the continuous-time survival literature, this involves the idea that there may be variability in individuals’ underlying (baseline) risk for an event that is not directly measurable, i.e., some individuals are more “prone” to an event than others.
Modeling approaches

- **Parametric:** Assume some underlying parametric distribution for \( \eta \) and maximize the likelihood.

- **Nonparametric:** Use a finite mixture, i.e., latent classes (\( C = 1, \ldots, K \)), to nonparametrically approximate the distribution of \( \eta \). (Long-term survivor model is a special case.)

- **Events:** 1\(^{st} \) arrest, 2\(^{nd} \) arrest, 3\(^{rd} \) arrest

- **Research question:** How do individuals differ with respect to underlying susceptibility to arrest and what predicts that individual-level variability?
Latent variable predictors

Indirectly measured risk factors for an event over time. Such a risk factors could be modeled as a categorical or continuous latent variables. These risk factors could be time-varying or time-invariant.

Event: First departure from full-time classroom teaching.
Predictor: School climate* measured by a set of indicators related to administration and student characteristics along with school resources. Time-varying since teachers may switch schools.
Research question: How does school climate affect if and when teachers leave the classroom?

*Could be teacher-specific or could use multilevel formulation if teachers are nested within schools.

Growth process: Teacher-rated aggression in grades 1-6.
Event: School removal in grades 7-12.
Research question: What is the relationship between aggression across grades 1-6 and the risk of school removal across grades 7-12?
Event occurrence represents an individual’s transition from one “state” (pre-event) to another “state” (post-event).

- States can be
  - Physical (e.g., living in a homeless shelter)
  - Psychological (e.g., depressed or healthy)
  - Social (e.g., married or divorced)
DTSA as LTA Extensions

- Multiple indicators of event status
- Dual process DTSA

Multiple indicators of event occurrence

- In some applications, event occurrence is inferred through indirect observation of the presence/absence of one or more symptoms that are used collectively (e.g., behavior checklist) to arrive at a “definitive” clinical diagnosis.
- In some settings, multiple measures of onset are used but obtain conflicting information, e.g., “have you ever had a drink?”, “at what age did you first drink?”, “number of drinks days in past month?”

Transition model:
Restrictions for $T=j$ to $T=j+1$ ($j \geq 2$):

<table>
<thead>
<tr>
<th>$\tau_{j,j+1}$</th>
<th>$e_{j+1} = 0$</th>
<th>$e_{j+1} = 1$</th>
<th>$e_{j+1} = 2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$e_j = 0$</td>
<td>$1 - P_h(j+1)$</td>
<td>$P_h(j+1)$</td>
<td>0</td>
</tr>
<tr>
<td>$e_j = 1$</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>$e_j = 2$</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
</tbody>
</table>
Quantifying error

- Symptom sensitivity: \( P(u_{pj} = 1 \mid e_j = 1) \)
- Symptom specificity: \( P(u_{pj} = 0 \mid e_j \neq 1) \)

The trouble with error

- Ignoring measurement error on event occurrence (and, thus, risk status) can result in either upward- or downward-biased hazard probability estimates.
- The impact of measurement error on the baseline hazard estimates depends on:
  - Number of symptoms
  - Sensitivity and specificity of each symptom
  - “True” baseline rate

Dual Process DTSA

Unlike time-varying predictors of events or competing risks, dual process DTSA models do not prioritize one event over another.

Events: High school dropout and onset of illicit drug use.

Research question: What is the relationship between the risk of high school dropout and the risk of illicit drug use over time?
Each “latent” class is a subgroup of adolescents that share the same age of onset, e.g., Class 3 consists of individuals that had their first real drink of alcohol at age 11. Class membership is known for all individuals with observed/reported onset age and partially known for right-censored individuals.
Onset-to-growth

At each time point there are two types of zeros:
1) Pre-onset zeros (individuals who have not initiated use).
2) Transient abstainers following onset.

Intra-individual change

For onset-to-growth, there are two process models
- Event history – describes the time-to-event process
  – Models the probability that an individual will transition from “at risk” for onset to initiating alcohol use.
  – Marks the beginning of the use trajectories, \( T_i=0 \)
- Post-onset use trajectories
  – Models the expected level of use across time as a function of time, conditional on initiation.

Inter-individual differences

- Event history – individual differences in the hazard of onset
- Growth trajectories – individual differences in growth trajectory parameters (e.g., intercept and slope growth factors)
Hazard of Alcohol Use Onset

Survival of Alcohol Use Onset

Model-estimated Levels of Use by Onset Age

Multi-faceted Longitudinal Process

- Timing of first lapse
- Use vs. non-use following lapse
- Risk and protective factors
- Frequency of use following first use, i.e., % days drinking
- Intensity of use following first use, e.g., drinks per drinking day
- Coping skills at intake were related to timing of first lapse and variability *within* post-lapse drinking classes.

- Alcohol Dependence Scores were related to timing of first lapse, variability *within* post-lapse drinking classes, and post-lapse drinking class membership.

- The timing of first lapse was related to post-lapse drinking class membership.
Exciting New Extensions

- Intersection with methods for intensive longitudinal data (e.g., micro-coded behavior observations of mother-child interactions)
- Causal effects of event timing on life course trajectories

Thank you 😊

Select References