SOC 7717 Event History Analysis and Sequence Analysis

Week 2: Event History Basics

Wen Fan

Spring 2019

Outline

Terms of Event History Analysis

Why Use Event History Analysis?

Censoring and Truncation

Describing the Distribution of Event Times

Approaches to Event History Analysis

Terms of Event History Analysis

Event history analysis

Event history analysis (sociology) = Survival analysis (epidemiology) = Failure-time analysis (engineering) = Reliability analysis (engineering) = Duration analysis (economics) = Hazard analysis (economics)

Collection of methods in which the aim is to analyze length of time until the occurrence of some event (*whether* and *when* events occur)

- Many different approaches
- All deal with censored data (esp. right-censored data)
- Durations are always positive and their distribution is often positively skewed

Examples of applications

- Health: death; disease onset; hospital stay; time to relapse
- Sociology: revolutions; social policies
- Demography: first birth; marriage; divorce; living in same house or area
- Economics: employment or unemployment; stock market crashes
- Education: leave full-time education; exit from teaching profession
- Engineering: equipment failures

Survival (time-to-event) data

Survival data are generated when the response measurement of interest is the time from a well-defined origin of measurement to occurrence of an event of interest

Three basic requirements define time-to-event measurements:

- Precise definition of occurrence of the event of interest
- Unambiguous origin for the measurement of time
- Agreed scale of measurement for time

Event

- A qualitative change that can be situated in time
  - Ideally, a change from one discrete state to another that occurs virtually instantaneously, e.g., death, marriage, promotion
- Can also talk about events with respect to quantitative variables if the change is sharp
  - E.g., an event defined as crosses a threshold of some quantitative variable (e.g., falling into poverty)
- You need to know *when* the change occurred

Event

In some studies, the event of interest (e.g. death) is bound to occur if we are able to follow-up each individual for a sufficient length of time

- However, whether or not the event of interest is inevitable has no consequence for the design, analysis, or interpretation of the study

In some studies the time-to-event (or survival probability) is of primary interest whereas in many social sciences studies we may be primarily interested in how various factors predict the event rates

Event

In many applications, each person can occupy only two possible states

- E.g., employed or unemployed, drinking or abstinent
- This is the assumption we assume until Week 8

In other applications, each individual can occupy three or more possible states

- E.g., being in school, dropping out, and graduation
- Competing risks

Beginning of time

A moment when everyone in the population occupies one, and only one, of the possible states

The goal is to "start the clock" when no one in the population has yet experienced the event but everyone is at least eligible to do so

Some possibilities:

- Birth
- The occurrence of a precipitating event: e.g., date of hospital release in a study of alcohol relapse
- An arbitrary start time that is unrelated to event occurrence: e.g., date of randomization

Metric for time

Units in which time is recorded: continuous or discrete

In generally, time should be recorded in the smallest possible units. However, there are situations in which we want to use discrete time

- Some events can occur only at discrete points in time (e.g., graduation dates)
- Many events are experienced in a discrete way
- Data collection constraints

Ties

Event history data

A longitudinal record of when events occurred for individuals or other units of analysis

- Prospective
- Retrospective (recall error, time-dependent covariates, truncation)

If the aim is to explain, the data should also contain information on possible explanatory variables. Some of these (e.g., sex) may be constant, while others (e.g., income) may vary over time

Why Use Event History Analysis?

A hypothetical question

Suppose we want to study recidivism among 500 inmates released from Massachusetts state prisons, followed for one year after their release

- Events: arrests
- Explanatory variables: financial aid, education, employment status

What are some potential models to use?

Potential methods

- **Method A**: dummy dependent variable (1 = arrest, 0 = no arrest). Logistic regression
  - Problems?

Potential methods

- **Method A**: dummy dependent variable (1 = arrest, 0 = no arrest). Logistic regression
  - Problems?
  - Wastes information
  - How to deal with employment status, which varies over time?
- **Method B**: dependent variable is length of time from release to first arrest. Poisson or negative binomial regression
  - Problems?

Potential methods

- **Method A**: dummy dependent variable (1 = arrest, 0 = no arrest). Logistic regression
  - Problems?
  - Wastes information
  - How to deal with employment status, which varies over time?
- **Method B**: dependent variable is length of time from release to first arrest. Poisson or negative binomial regression
  - Problems?
  - What about cases with no arrests (censoring)?
  - How to include time-varying explanatory variables?

Two central problems

- How to deal with censoring (i.e., combine information for those who did and did not experience events)?
- How to incorporate explanatory variables which vary over time?

Censoring and Truncation

Right censoring

- Suppose we have a random variable T
- We say that a particular observation of T is right censored if all we know about T is that it is greater than some constant c
  - E.g., suppose we have a sample of women interviewed at age 30, and the event of interest is first marriage. Let T be the age at marriage. For women still unmarried, we know only that $T > 30$

Right censoring

Right censoring is the most common form of censoring in social science, so let's think about some common reasons for right censoring:

- Termination of the study before the event occurs
- Death due to a cause not considered to be the event of interest
- Loss to follow-up
- Withdraw permanently

Left censoring

- A particular observation is called left censored if all we know about T is that it is less than some constant c
    - E.g., suppose we do a prospective study among a cohort of women aged 25 to determine their age at first marriage. But in the initial interview, we only ask if these women are currently married, not the year. For those already married, we know only that $T < 25$

Interval censoring

- A particular observation is interval censored if all we know about T is that it is between two numbers a and b
    - E.g., suppose we interview a sample of women in 2015 and 2018 and we ask for marital status in each year. Those who are unmarried in 2015 and married in 2018 have marriage times that are interval censored

Types of right censoring: fixed vs. random

- Fixed censoring: for each case in the sample ($i = 1, ..., n$), there is a number c_i (determined in advance by the study design) such that if $T_i \leq c_i$, then T_i is observed, while if $T_i > c_i$, the case is censored
    - Special case: $c_i = c$ for all i (i.e., "singly right censored")
    - E.g., recidivism study, released inmates are followed for one year after release. T is the number of months from release to first arrest and $c_i = 12$ for all cases

Types of right censoring: fixed vs. random

- Random censoring: same as fixed, except that the c_i's are random variables rather than being determined by the study design. Occurs when individuals drop out, die, or are otherwise lost to follow-up
  - E.g., sample consists of a cohort of entering sociology graduate students at Boston College. T is length of time from entry to receipt of Ph.D. Follow-up for seven years
  - Those still registered but haven't received degree at the end of the seventh year are censored by a fixed mechanism. But many others will be censored at earlier times because of drop out (random censoring)

Types of right censoring: informative vs. non-informative

If censoring is random (as opposed to fixed),

- Noninformative censoring: conditional on the explanatory variables, the fact that an individual is censored at time t does not give any information about the individual's hazard at time t. That is, individuals are not censored because they are at higher or lower risk of an event
- One way to think of this is that, conditional on the values of any explanatory variables, the individuals censored at time t should be a random sample of the individuals at risk at time t

Informative censoring

- When censoring is associated with risk of an event, this is known as informative censoring, and standard methods of analysis will result in biased estimates
- Common methods for controlling for informative censoring are to stratify or condition on those explanatory factors on which censoring depends
- Generally, the consequences of informative censoring are difficult to predict or gauge, but can be investigated by sensitivity analysis

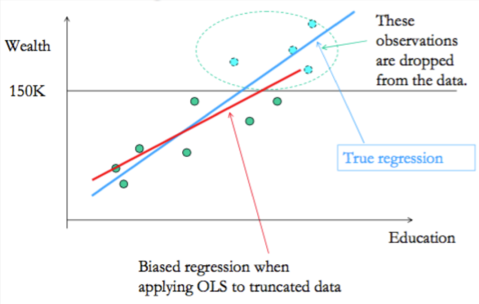
Truncation

Reasons for truncation:

- Truncation by survey design
  - E.g., studies of poverty. By design, families whose incomes are greater than that threshold are dropped from the sample
- Incidental truncation: it is the people's decision, not the survey design, that determines the sample selection
  - E.g., only those who are working have wage information

Why truncation matters?

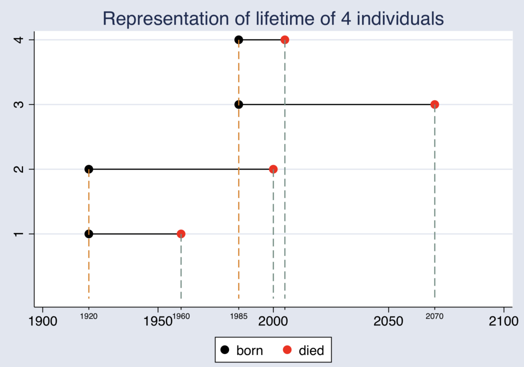
If we estimate only on the sample for which we have information, coefficients are biased



Censoring and truncation

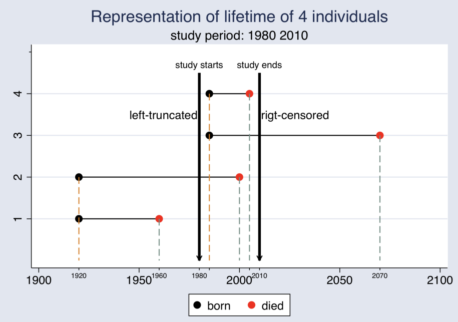
- Censoring: everyone is included in our sample, but for some individuals, the exact value of Y is unknown; we only know it to be within a range
  - The outcome may be censored, but you can include the censored observations in the regression
- Truncation: we know nothing about those below or above some threshold on Y; they are excluded from our sample
  - A subset of observations are dropped; thus, the truncated data are not available for the regression

Exercise: a toy example



Exercise: a toy example

Let's assume that we conducted a longitudinal study that went from 1980 to 2010:



Exercise: a toy example

What would our data look like?

id	born	study_beg	enter	last_observed	died

Note

Standard methods for survival analysis assume that all censored data are right censored and we will assume that this is the case

Special methods are required for analyzing left censored, interval censored, or truncated data, which will not be covered in this course

Describing the Distribution of Event Times

Terminology

- A ratio is the result of dividing one quantity by another, with the numerator and the denominator two separate and distinct quantities
- A proportion is a type of ratio in which the numerator is included in the denominator
- A rate is a measure of change in one quantity per unit of another quantity. Rates typically measure events per unit time
    - The “survival rate” of a group of people over a specified time period is therefore not really a rate, but a proportion

Distributions of event times

In all approaches to event history analysis, the event time T is regarded as random or stochastic. Accordingly, we can describe it in ways that are standard for random variables

- Cumulative distribution function (cdf): $F(t) = Pr(T \leq t)$
- Survival function: $S(t) = 1 - F(t)$, indicating the probability of "surviving" past time t
- Probability density function (pdf): $f(t) = \frac{\partial F(t)}{\partial t} = -\frac{\partial S(t)}{\partial t}$
- Hazard function: $h(t) = \lim_{s \to t} \frac{Pr(t<T<s|T \geq t)}{s-t}$

Relationships between functions

Equivalences between functions:

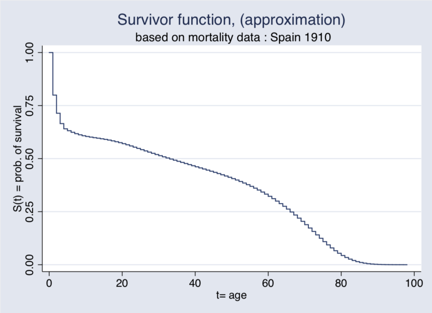
$$h(t) = \frac{f(t)}{1 - F(t)} = \frac{f(t)}{S(t)}$$

$$f(t) = h(t)\exp(-\int_0^t h(u)du)$$

$$F(t) = 1 - \exp(-\int_0^t h(u)du)$$

Survival function S(t)

The survivor function, $S(t)$, gives the probability of surviving until at least time t



Survival function S(t)

- A nonincreasing function with a value 1 at the time origin and a value 0 as t approaches infinity
- A function that depends on t
- The survivor function evaluated at a specific value of t is often referred to as the "survival rate", for example, the "5-year survival rate"
  - As emphasized above, it is actually "survival proportion"

Survival function S(t)

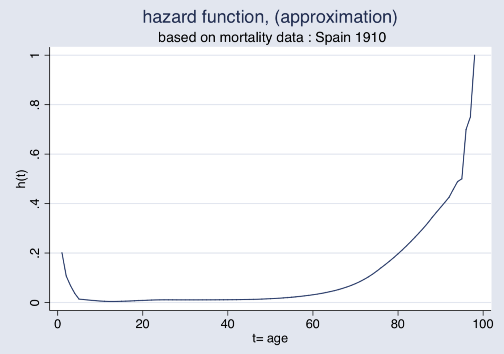
- Nonparametric methods for estimating $S(t)$ (to be discussed next week) involve estimating the survival proportion at discrete values of t and then interpolating these to obtain an estimate of $S(t)$
- In studies where incidence is the outcome, we often present the cumulative incidence, given by $1 - S(t)$, rather than $S(t)$
- It is generally difficult to determine the essence of the failure/survival pattern, and even more difficult to compare it between groups, simply by studying plots of the survivor function $\longrightarrow$ hazard function $h(t)$

Hazard function h(t)

$$h(t) = \lim_{s \to t} \frac{Pr(t<T<s|T \geq t)}{s-t}$$

- Aka: rate, hazard, intensity, hazard rate, force of mortality
- A way of mathematically expressing the intuitive notion of the risk of event occurrence: the probability that an event occurs between t and s given that an event has not already occurred
- Sometimes referred to as “instantaneous probability of event occurrence”
  - Not really appropriate because $h(t)$ is a rate, not a probability, so it has no upper bound

Hazard function h(t)



Hazard function h(t)

- Like probability, hazard is never directly observed
- It is a rate: number of events per interval of time
- If $h(t)$ is a constant c, then c is the expected number of events in an interval that is one time-unit long
  - E.g., $h(t)$ = .78 for all t. Then we expect 0.78 events per unit time

Hazard function h(t)

Example: We observe 10,000 individuals for a period of one year, during which the hazard is constant. Two hundred of them die during this year

- Let U be the total amount of time (in years) that all individuals are observed. For all those who don't die, the contribution to U is 9,800. For those who do die, the contribution to U is the length of time from the beginning of the year to the time of death
- Then an optimal estimate of the hazard is 200/U

Hazard function h(t)

If $h(t)$ is constant over t, then the expected value of T (i.e., length of time until the event occurs) is $E(T) = \frac{1}{h(t)}$

- E.g. if $h(t)$ = 0.78 for all t and time is measured in years, then 1/0.78 = 1.28 years is expected length of time until an event occurs

Note that we usually assume the hazard be a function of t so that the instantaneous risk can vary with time

Hazard function h(t)

- In contrast to the survivor function, which describes the probability of not failing before time t, the hazard function focuses on the failure rate at time t among those individuals who are alive at time t
- That is, a lower value for $h(t)$ implies a higher value for $S(t)$ and vice-versa

Expectation of life

- Calculated as the area under the survivor function
- Extrapolation techniques can be used if the survivor function does not reach zero while the respondents are under follow-up
- By comparing the expectation of life of the patients to the expectation of life of a comparable group from the general population, it is possible to estimate the "proportion of expected life lost"

Median survival time

- The time beyond which 50% of the individuals in the population are expected to survive
- Estimated by extrapolation if the cumulative observed survival proportion does not sink below 0.5 during the period the patients are under follow-up

Approaches to Event History Analysis

Different approaches

- Nonparametric: life table, Kaplan-Meier estimators (Week 3)
- Parametric: exponential regression, log-normal regression, etc. (Week 5)
- Semi-parametric: Cox proportional hazards regression (Weeks 6-7)
- Discrete-time methods (Week 7)