SOC 7717 Event History Analysis and Sequence Analysis

Week 9 Lecture 1: Discrete-time Event History Analysis

Wen Fan

Spring 2019

Outline

Discrete-time Data

Discrete-time data

In social research, event history data are usually collected

- Retrospectively in a cross-sectional survey, where dates are recorded to the nearest month or year, or
- Prospectively in waves of a panel study (e.g. annually)

Both give rise to discretely-measured durations

Also interval-censored because we only know that an event occurred at some point during an interval of time

Discrete-time data

More broadly, events can occur at any time but measurement of time is not precise

- No matter how small the units are, we still measure time in discrete units
- Sometimes the units are large (e.g., months, years, or decades) relative to the total period of observation and the rate of event occurrence $\longrightarrow$ continuous time methods may not be appropriate
- A good indication of the need for discrete-time methods is the presence of substantial numbers of ties

General analytic steps for discrete-time data

1. Break each individual's event history into a set of distinct observations (e.g., person-years), one for each unit of time until censoring or an event occurs
2. For each of these observations, code the outcome as 1 if an event occurs during that time unit, otherwise 0. Explanatory variables take on whatever value occurs during that time unit
3. Pool these observations and estimate a binary logit regression model (`logistic`) or a complementary log-log model (`cloglog`) by maximum likelihood

Data preparation for discrete-time analysis

- We must first restructure the data into long form
- We expand the event times and censoring indicator to a sequence of binary responses y_{ti}, where y_{ti} indicates whether an event has occurred in the time interval $[t, t+1)$

Discrete-time Logit Model

Discrete-time hazard function

Let's use p_{ti} to denote the probability that individual i has an event during interval t, given that no event has occurred before the start of t

$$p_{ti} = \Pr(y_{ti} = 1 | y_{t-1,i} = 0)$$

p_{ti}, referred to as the discrete-time hazard function, is a discrete-time approximation to the continuous-time hazard function $h_i(t)$

Discrete-time logit model

After expanding the data, we fit a binary logit model to y_{ti}:

$$\log(\frac{p_{ti}}{1 - p_{ti}}) = \mathbf{A}D_{ti} + \mathbf{B}\mathbf{X}_{ti}$$

- p_{ti} is the probability of an event during interval t (constrained by $0 \leq p_{ti} \leq 1$)
- D_{ti} is a vector of functions of the cumulative duration by interval t with coefficients $\mathbf{A}$
- $\mathbf{X_{ti}}$ is a vector of covariates (time-varying or constant over time) with coefficients $\mathbf{B}$

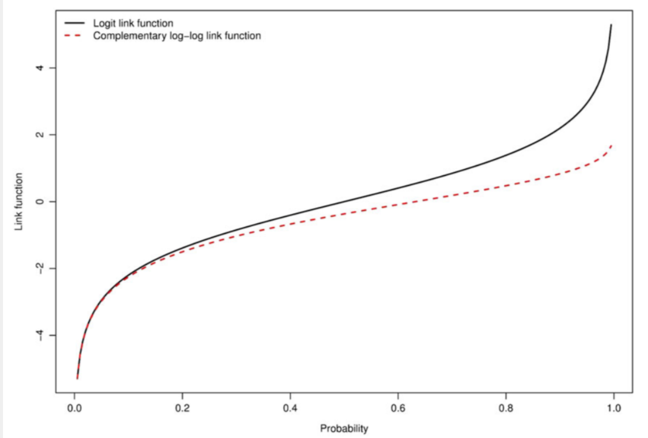
Complementary log-log model

If the data were really generated by a proportional hazards model in continuous time, the correct functional form is the complementary log-log:

$$\log[-\log(1 - p_{ti})] = \mathbf{A}D_{ti} + \mathbf{B}\mathbf{X}_{ti}$$

- The coefficients in this model are exactly equivalent to the coefficients in the underlying Cox model
- Unlike the logit, this model is invariant to the interval length
- The function is asymmetrical $\longrightarrow$ always set up the model to predict the probability of an event rather than a non-event

Shapes of logit and complementary log-log functions



D_{ti}: Time-dependency of the hazard

Changes in p_{ti} with t are captured in the model by $\mathbf{A}D_{ti}$, the baseline hazard function

D_{ti} has to be specified. Options include:

- Step function: $\mathbf{A}D_{ti} = \alpha_1 D_1 + \alpha_2 D_2 + ... + \alpha_q D_q$, where $D_1, ..., D_q$ are dummies for time intervals $t = 1, ..., q$ and q is the maximum observed event time
  - If q is large, categories may be grouped to give a piecewise constant hazard model
- Polynomial up to order q: $\mathbf{A}D_{ti} = \alpha_0 + \alpha_1 t + ... + \alpha_q t^q$
- ...

Estimation methods

Maximum likelihood: easily applied and extremely flexible, especially for time-varying explanatory variables

- Allows for great flexibility in specifying the time function. In contrast to continuous-time models, time is just another variable on the right-hand side
- Can have more than one time scale

Non-proportional hazards

- So far we have assumed that the effects of X are the same for all values of t
- It is straightforward to relax this assumption in a discrete-time model by including interactions between X and t in the model
- Test for non-proportionality by testing the null hypothesis that the coefficients of the interactions between X and t are all equal to zero

Discrete-time Poisson Model

Discrete-time Poisson model

$$\log(\lambda) = \beta_0 + \beta X_{ti}$$

β indicates the effect per unit of X, on the log rate scale

In Stata:

- `poisson y x, exposure()`
- Poisson regression can also be performed using the `streg` command with `dist(exponential)`. This is preferable when the data have been `stsplit`

Discrete-time Poisson model

- Estimated using the method of maximum likelihood
- Confidence intervals are constructed by assuming the estimated regression parameters are normally distributed
- The confidence limits for the incidence rate ratio (IRR) are simply the exponentiated limits of the log IRR
  - As such, the CI for the IRR is not symmetric around the point estimate

Comparison of Cox and Discrete-Time Regressions

Similarity

The methods are very similar; the basic formulation of both models is

$$\log(rate) = \beta X$$

- In both cases, the β parameters are interpreted as log rate ratios
- Both models are multiplicative (i.e., both assume proportional hazards)

Similarity

When will Cox and discrete-time estimates be similar?

- A discrete-time model with a complementary log-log link, $\log(-\log(1-p_t))$, is an approximation to the Cox proportional hazards model, so the coefficients are directly comparable
- In general, Cox and logit estimates will get closer as the hazard function becomes smaller
  - The discrete-time hazard will get smaller as the width of the time intervals become smaller

Differences

In discrete-time regression,

- Follow-up time is classified into bands and a separate rate parameter is estimated for each band, thereby allowing for the possibility that the rate is changing with time
- The rate is assumed constant within each band
- We are not forced to choose a single scale for "time"

Differences

In Cox regression,

- We essentially choose bands of infinitesimal width; each band is so narrow that it includes only a single event
- We do not estimate the baseline rates within each time band; instead, we estimate the relative rates for the different levels of the covariates
- Cox regression is more efficient in this respect if we have a small study (few events)