

Discrete Time Event History

Nathaniel Beck (with a little help from his friends, particularly
Jonathan Katz)

Department of Politics, NYU, New York, NY 10012, nathaniel.beck@nyu.edu

Q2 - Spring 2009

1

Event History

- Much BTSCS data has long strings on 0's with few 1's
- E.g. conflict data (perhaps 3% of obs are 1's, with generous def of 1
- looks like event history data
- with each 1 marking a failure
- and the time between 1's, that is, the number of 0's, being the time until or between failures
- Make sure the number of 1's is not large, so we have a good distribution of failure times. item We could thus simply convert the binary TSCS data into event history data
- and use standard (Cox or parametric)

2

Helpful event history ideas

- Thinking about binary TSCS data as event history data helps
- The simple probit/logit approach is equivalent to the assumption of no duration dependence in event history analysis.
- normally we test for duration dependence,
- While it looks like we have $N \times T$ binary TSCS observations
- this is the same as N duration observations
- While we think of probit/logit as having troubles with rare events
- such rare events are the lifeblood of event history analysis
- Binary TSCS data in event form may have more than one event per unit.

3

Onset vs. Incidence, repeated events

- In the event history approach, we model strings of zeros which end with a 1
- that is, the probability of a transition from 0 to 1,
- or what is known in the medical world as ONSET (of a disease).
- We are NOT modelling the length of strings of 1's.
- The total proportion of 1's is called INCIDENCE. (Proportion of all people having the disease.)
- We could model length of time of string of 1's (spells of disease, war)
- THIS TURNS OUT TO BE A CRITICAL ISSUE (return to below)
- In the probit/logit setup, we assume that second and subsequent events can be modeled just like first events
- events history modelers realize that life is more complex
- Solution???? Kludge??? Counter for how many prior events

4

- Sticking with one event per unit for now
- Can model the time until an event in binary TSCS data by a discrete time duration model.
- Assume time measured in equal discrete intervals, $0, 1, \dots, t, \dots$ (years)
- We only observe whether someone dies in the interval $(t - 1, t]$, (open on the left, closed on the right)
- and will assume this is a death at t
- Then we need discrete time analogues of the survivor and hazard function.
- The survivor function will simply be a step function, with steps at $1, 2, \dots, t, \dots$
- Let y_i be the duration for the i 'th unit;
- y_i is a discrete random variable, with support at the positive integers.

5

Discrete time maths

- Using standard notation we have

$$\begin{aligned} S(t) &= P(y > t) \\ &= P(y > t | y > t - 1)P(y > t - 1) \\ &= \prod_{i=0}^{t-1} P(y > t - i | y > t - i - 1) \end{aligned} \quad (1)$$

where $S(0) = P(y > 0) = 1$. We still have

$$F(t) = 1 - S(t). \quad (2)$$

- Since F is discrete, we have an associated discrete density with support on the positive integers,

$$f(t) = F(t) - F(t - 1). \quad (3)$$

Here the density is a probability; it is the unconditional probability of dying in the interval $(t - 1, t]$.

6

- Define the discrete hazard analogously to the continuous time hazard
- though simpler, since is now just a conditional probability
- $h(t)$ is the hazard of dying at time t (or in the interval $(t - 1, t]$ given that one survived until $t - 1$
- that is, probability of death in that interval given alive at start of interval

$$h(t) = \frac{f(t)}{S(t-1)}. \quad (4)$$

Since $1 - h(t)$ is the conditional probability of surviving at t given survival through $t - 1$, substituting in Equation 1, we get

$$S(t) = \prod_{i=0}^{t-1} [1 - h(t - i)] \quad (5)$$

7

Estimation of discrete duration models via logit

- Estimating a BTSCS with dependent variable $y_{i,t}$ being whether unit i failed in the interval $(t - 1, t]$
- (by probit, logit or any other binary model, will use “logit” as generic)
- as a function of covariates $\mathbf{x}_{i,t}$
- we are estimating a model for $h(t)$
- If the dependent variable is scored as 1 for non-failure, then we have a model for $1 - h(t)$
- Estimating via ordinary logit is assuming the hazard rate is time invariant (that is,

$$h_{i,t} = h(\mathbf{x}_{i,t}) \quad (6)$$

- To allow for duration dependence estimate we would need to estimate a binary model (with $y_{i,t}$ being one for the failure of unit i in the interval $(t - 1, t]$ which has

$$h_{i,t} = h_t(\mathbf{x}_{i,t}). \quad (7)$$

8

Separate time counter

- Compromise to allow for different intercepts at each time point

$$h_{i,t} = a_t + h(\mathbf{x}_{i,t}) \quad (8)$$

- which would be estimated by putting in a period dummy in the logit.
- Or could use any $s(t)$ one liked, if flexible
- Remember, *the time variable is time since the last “event,” not the particular period of the observation.*
- Sometimes the time dummies indicate that we don't need to include time in the specification (using standard tests on the coefficients of all the time variables).
- At that point we can assume no duration dependence and use ordinary logit.

9

A more formal derivation

- Start with a continuous time Cox proportional hazards model, so

$$h_i(t) = h_0(t)e^{\mathbf{x}_{i,t}\beta}. \quad (9)$$

- Letting $S(t)$ be the probability of surviving beyond t , by the math of hazard rates we have

$$S(t) = \exp\left(-\int_0^t h(\tau)d\tau\right). \quad (10)$$

- Only observe whether or not an event occurred between time t_{k-1} and t_k

- so model $P(y_{i,t_k} = 1)$

$$\begin{aligned}
 P(y_{i,t_k} = 1) &= 1 - \exp\left(-\int_{t_{k-1}}^{t_k} h_i(\tau) d\tau\right) \\
 &= 1 - \exp\left(-\int_{t_{k-1}}^{t_k} e^{\mathbf{x}_{i,t_k}\beta} h_0(\tau) d\tau\right) \\
 &= 1 - \exp\left(-e^{\mathbf{x}_{i,t_k}\beta} \int_{t_{k-1}}^{t_k} h_0(\tau) d\tau\right)
 \end{aligned}$$

11

Even more maths

- Since the baseline hazard is unspecified
- Can just treat the integral of the baseline hazard as an unknown constant
- Defining

$$\begin{aligned}
 \alpha_{t_k} &= \int_{t_{k-1}}^{t_k} h_0(\tau) d\tau \text{ and} \\
 \kappa_{t_k} &= \log(\alpha_{t_k})
 \end{aligned}$$

- we then have

$$\begin{aligned}
 P(y_{i,t_k} = 1) &= 1 - \exp\left(-e^{\mathbf{x}_{i,t_k}\beta} \alpha_{t_k}\right) \\
 &= 1 - \exp\left(-e^{\mathbf{x}_{i,t_k}\beta + \kappa_{t_k}}\right)
 \end{aligned}$$

- This is exactly a binary dependent variable model with a cloglog link and the κ (time dummy) terms added.
- There is almost no difference in practice between estimated a logit model and a cloglog model

12

- Thinking about the war data as event history data leads to thinking about other issues.
- Dyads can fight a number of wars.
- Durations of second events may follow different process than for first events
- This is difficult to model
- One solution is to add a variable to the hazard function which counts the number of previous failures
- Another issue is modeling onset vs. incidence
- To understand we need a detour