SOC 7717 Event History Analysis and Sequence Analysis

Week 10 Lecture 1: Case Study and Review of Event History Analysis

Wen Fan

Spring 2019

Outline

Case Study

Case study: summary

- Research questions
  - Youth violent victimization $\longrightarrow$ timing of dating debut and first union formation
  - Moderation by age or gender
- Theoretical approach
  - Life course: timing and sequence
  - Competing hypotheses: "rejection sensitive" vs."overinvestment"
- Data: Add Health
- Methods: Cox regression, competing risk model
- Findings
  - Victims begin dating sooner and progress more quickly from dating to first unions than do non-victims
  - Age-graded patterns but no gender difference

Case study

- Measures
  - Note how they defined their "event" (pp. 1248–1250)
    - Shifts in definition across waves (pp. 1249–1250)
  - Key predictor: youth violent victimization (Sebastian; Xiyao)
  - Choice of covariates and time order (Ryan; Sebastian)
- Methods
  - Left-censoring: 40% (p. 1249) (Ryan; Xiyao)
    - Note how they dealt with the problem and justified their decision
  - Proportional hazards assumption (Sebastian)
  - Tied data (Ryan)
  - Model choice (Xiyao)
- Tables/Figures
  - Table 2: by key predictor
  - Table 3: descriptive table (%event and event time) (why Wilcoxon test?)
  - Figures: better present hazards with CIs

Miscellaneous Topics

How to Choose a Method?

- Make Cox models your default method
  - More robust than parametric models; allows time-dependent covariates; allows for temporary exit from the risk set; widely used and understood
- Is the sample large with heavily tied event times? $\longrightarrow$ discrete-time event history models
- Want to study the shape of the hazard function (e.g., constant? direction of change? rate of change?) $\longrightarrow$ parametric models or discrete-time event history models
- Want to generate predicted event times or survival probabilities? $\longrightarrow$ parametric models

Multicollinearity

Everything you know about multicollinearity for ordinary regression models carries over to hazard regression models

To diagnose, use a multiple regression program to do a preliminary run and request multicollinearity diagnostics (`vif`)

Repeated Events

Repeated Events

Most events of interest in the social sciences are, in principle, repeatable:

- births
- marriages
- job changes
- promotions
- visits to a physician
- residential mobility
- arrests

Two commonly-used approaches

1. Do a separate analysis for each successive event, e.g., birth
   - Do one analysis for marriage to first birth
   - Do a second analysis for 1st birth to 2nd birth (among those who had a 1st birth)
   - ...

Good approach if process changes after each event (e.g., if effect of some variable is stronger for first birth than for second birth)

Tedious and wasteful of information if process is invariant

Numbers of cases usually declines substantially with succeeding events

Two commonly-used approaches

2. Treat each successive interval as a distinct observation, and pool all such intervals

E.g., a 5-year observation period:

|———×———————×————×————————|

0 5

Three events, four intervals (last one censored)

More observations, more statistical power

What's wrong with this method?

Two commonly-used approaches

Problem 1: Assumes that the social process is invariant from one interval to the next

- Coefficients are the same
- Probability distribution is the same

Using Cox with stratification, it's possible to relax the second restriction while preserving the first (i.e., a different arbitrary function for each successive interval):

First interval: $\log h(t) = \alpha_1(t) + \beta_1 X_1$

Second interval: $\log h(t) = \alpha_2(t) + \beta_1 X_1$

Two commonly-used approaches

Problem 2: Assumes that multiple intervals for each individual are independent, conditional on the explanatory variables

- The assumption of (conditional) independence is probably always violated to some degree $\longrightarrow$ standard errors can be seriously biased downward

How to test the independence assumption?

- Estimate a model for 2nd intervals with length of 1st interval as one of the explanatory variables $\longrightarrow$ a significant coefficient indicates dependence

More advanced approaches: marginal methods

Marginal methods or population-average methods ignore the possible downward bias in coefficient estimates, and concentrate instead on getting better standard error estimates

The coefficients are unchanged from conventional Cox regression. The standard errors, z-statistics, and p-values are adjusted for clustering (p-values usually get larger) (`cluster()`)

More advanced approaches: conditional methods

Conditional methods or subject-specific methods assume that the dependence is created by unobserved heterogeneity, and they correct for bias in both coefficient estimates and standard error estimates

Basic model is:

$$\log h_{ij}(t) = \alpha_t + \mathbf{BX}_{ij} + \epsilon_i$$

where $h_{ij}(t)$ is the hazard for the j-th event for the i-th individual. The ϵ_i's i represents unobserved heterogeneity that is specific to an individual

More advanced approaches: conditional methods (RE)

Random effects models: the unobserved heterogeneity term is treated as a random variable with a specified probability distribution

Disadvantages:

- Sensitivity to assumed distribution
- Heterogeneity term is usually assumed to be independent of the measured covariates

In Stata, random-effects parametric models can be estimated with `streg` with the `frailty() shared()` option or `stcox` with the `shared()` option

More advanced approaches: conditional methods (FE)

Fixed effects models: the unobserved heterogeneity between individuals is treated as a set of fixed constants

- Advantage
  - Unobserved heterogeneity can be correlated with measured variables that vary over time
  - Much weaker distributional assumptions than random effects
- Disadvantage
  - Can only include covariates which vary across spells; less efficient
  - Throws away a lot of information on ranks

More advanced approaches: conditional methods (FE)

Reformulate the model as proportional hazards model with an arbitrary function of time for each individual:

$$\log h_{ij}(t) = \alpha_i(t) + \beta X_{ij}(t)$$

Run Cox models with stratification on individuals (`stcox` with the `strata()` option)

Repeated Events for Discrete-Time Data

Can use the same approach as with continuous-time models

- Population-averaged methods
  - Use `cluster` option on `logistic` to adjust standard errors without changing coefficient estimates
  - Use generalized estimating equations (GEE) with `xtgee`. Produces more efficient estimates with robust standard errors
- Subject-specific methods
  - Estimate random-effects logit models with `xtlogit`
  - Estimate fixed-effects logit models with `xtlogit` or `clogit`

Origin time for repeated events

When events are repeated, most common approach is to "reset the clock to 0" each time an event occurs

It's also possible to define time in terms of a single origin point for all intervals, e.g., date of diagnosis or date of hire

- In other words, each successive spell for a single individual starts at a different time
  - » Most Cox regression programs assume that all spells start at time 0, which implies that everyone is in the risk set at the beginning
- This can be handled with the `time0` option in `stset`

Multiple Kinds of Events

Multiple kinds of events

We have assumed to this point that all events in a given analysis are indistinguishable, e.g., all deaths are the same, all job changes are the same, etc.

Many instances where we may want to make a distinction:

- May be completely inappropriate to lump different kinds of events together (e.g., firings and quittings)
- Even when appropriate, we may want a more refined analysis (e.g., arrests for violent vs. nonviolent crimes)

There are multiple kinds of multiple kinds of events $\longrightarrow$ much confusion in the literature

Classification

1. Parallel processes. The occurrence of each event type follows a distinct process (may involve different explanatory variables, different coefficients, or different functional forms)

Competing risks: the occurrence of one event type removes the individual from risk of the other event types:

- Death: cancer vs. heart disease vs. stroke vs. all others
- Marital dissolution: death vs. divorce
- Job termination: quitting vs. firing
- Changes of government: constitutional vs. coup d'etat

Classification

2. Conditional processes. The occurrence and timing of events is determined by one process. Given that an event occurs, a different mechanism determines which kind it is, e.g.,

- Event is purchase of a car and cars are distinguished by manufacturer
- Event is marriage and distinction is between civil and religious ceremony

Analysis:

- Use regular methods for timing of event
- Among events, use logit analysis (binomial or multinomial) to estimate models for type of event

Competing risks

Example: 5 kinds of deaths: cancer, heart disease, stroke, accident, all other

Let $j = 1, ..., 5$ index the different kinds of events. Let J be a random variable denoting which event type actually occurred

Competing risks

Define a type-specific (cause-specific) hazard function. Let $P_j(t, s)$ be the probability that a person dies from cause j in the interval (t, s) given that the individual was still alive at time t, i.e.,

$$P_j(t, s) = \Pr(T < s, J = j | T \geq t)$$

Then

$$h_j(t) = \lim_{s \to t} \frac{P_j(t, s)}{s - t}$$

Competing risks

For each $h_j(t)$ we can now specify whatever model is appropriate for dependence on explanatory variables

- The likelihood function for all competing risks can be factored into separate likelihood functions for each event type
- Each of these likelihood functions treats the occurrence of other events as if the individual was censored at the time of occurrence

Time Dependence and Heterogeneity

Time dependence

The general proportional hazards model treats the dependence of the hazard on time as a nuisance function:

$$\log h(t) = \alpha(t) + \beta X$$

Thus, we cannot estimate or test hypotheses about the form of this dependence

Time dependence

This is often an important research question, however

- Principle of cumulative inertia holds that the longer a person is in a given state, the less likely s/he is to leave that state
- "Liability of newness" hypothesis states that new organizations are more prone to dissolution than older organizations

Time dependence

- To test such hypotheses, we need a parametric modeling approach
  - E.g., estimate a Weibull model and determine whether the coefficient for $\log t$ is positive or negative
- But an unstated assumption is that two individuals with the same values of the X-variables at the same point in time will have exactly the same hazards
  - Unlikely to be the case because the measured X-variables almost never exhaust the differences between individuals
- Unobserved heterogeneity tends to produce empirically declining hazard functions
- Why?

Time dependence and heterogeneity

Example: Two homogeneous populations, $h_1(t) = \mu_1$, $h_2(t) = \mu_2$. But we don't know who is in which population; we can only observe their mixture

- Individuals with high hazards, die early and get removed from the risk set
- As time goes by, the remaining sample consists increasingly of low-risk individuals
- There's almost always some heterogeneity, so there's almost always some tendency for observed hazard rates to decline with time even when they are not declining for any individual
  - Implication: if the observed hazard rate is increasing with time, the hazard must truly be increasing for at least some proportion of the sample

Frailty models

Some attempts have been made to formulate models which allow for unobserved heterogeneity, e.g., an extended Weibull model

$$\log h(t) = \alpha + \beta_1 X_1 + \beta_2 X_2 + \beta_3 \log(u)$$

where u is an unobserved random disturbance term ("frailty")

Frailty models

Estimation is computationally difficult

- Results tend to be unstable and depend heavily on the postulated distribution of u or the functional form for dependence on time
- Not possible to effectively distinguish among different models

Heterogeneity

Does unmeasured heterogeneity bias estimates of the effects of other variables?

- If unmeasured variables are correlated with the measured variables, that can lead to severe bias. But that's true in ordinary regression

What if the u-term is independent of the X's?

- When there is censoring, all models yield biased estimates; coefficients are attenuated toward 0. But standard errors and hypothesis tests are still valid. This is also true in logit analysis
- When there is no censoring, exponential and Gamma models are not biased by unobserved heterogeneity