Event history models

Silje Synnøve Lyder Hermansen

2025-04-23

Where are we?

GLMs

Social scientists are often concerned with human behavior

- these are often "events"
  - a choice: the actor is doing something
  - a treatment: something occurs to the actor
- events are discrete, while linear models assume a continuous and outcome
  - theorize a latent variable that is continous: a "propensity" that translates to observable events
  - recode events to something continous and use probability distribution to link events to propensity (probability)

$\Rightarrow$ *the domain of Generalized Linear Models*

Different models of events

Different models of events focus on different aspects

- binomial logistic regression:
  - we have a "trial" and a success/error
  - covariates at the trial level
  - focus on whether the event happened
- event count models (e.g. poisson regression)
  - we have a window of opportunity ("exposure") and a number of events
  - covariates at the exposure level
  - focus on the number of events
- event history models (e.g. cox proportional hazard)
  - we have a duration ("spell") and an event/non-event at the end of the period
  - covariates at the spell level
  - we focus on the time between events

$\Rightarrow$ *sometimes we can pick any of these models*

Example: political violence

Example: political violence

Nanes (2017) "Political Violence Cycles: Electoral Incentives and the Provision of Counterterrorism"

- counter-terrorism is a signal to voters that election-seeking office holders care about voter security
  - study of Israeli checkpoints and attacks on Palestinians as a function of electoral cycle
- data generating process:
  - Prime Minister / cabinet members decide on a violent attack
- three potential operationalizations of the outcome:
  - a decision to kill
  - number of killed in a day
  - time between decisions

Event history models

Time

Political science is full of phenomena that involve time

- unemployment, cabinets (governments), war, peace, negotiations. . .
- they contain two components:
  - an event (a binary outcome)
  - a duration (a "spell")

$\Rightarrow$ *glass half full/half empty situation*

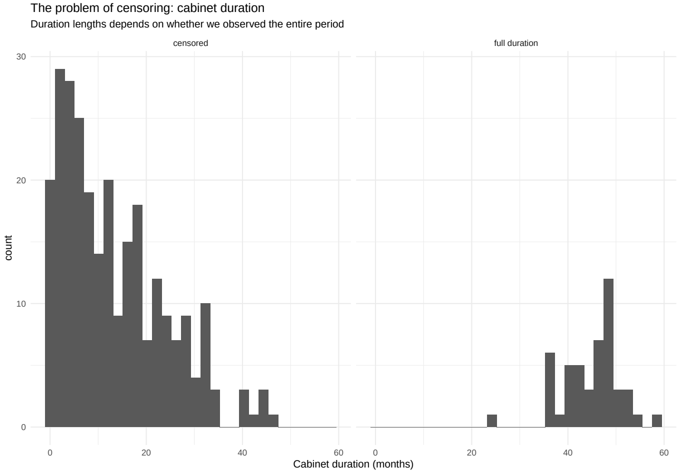
Class half-full? Or empty?

- **opportunity:** duration may be a substantive measure of its own
  - e.g. ability for a cabinet to stay in office
- **constraint:** observations are censored (i.e. we don't know when the spell ends)
  - we can't truncate them (code them with max observed length): bias the slope (β)
  - we can't remove them: bias the sample

$\Rightarrow$ *model them as such*

Censoring: two sets of observations

Censoring means that we don't observe the entire length of some spells



Choices and priorities

Choices and priorities

Event history models require us to make quite a few choices

- information leveraged in the **outcome** (event/duration)
- **unit of analysis** (spell, spell + change in x, TSCS)
  - as a consequence: the nesting/correlation between repeated measures
- **functional form** (survival or hazard; parametric/estimated or semi-parametric/empirical)
- **ties** what to do when many observations experience the same event?

Data structures: unit of analysis

Data structures: unit of analysis

Event history models often end up with complicated data structures and a host of “outcome” variables

- division between continuous time models and discrete time models is blurry
- unit of analysis:
  - spell-level: as simple as it gets (duration + event/no event)
  - spell-level + time varying covariates
  - fixed time period

$\Rightarrow$ *Think ahead, because you will be doing data-wrangling*

Spell-level: focus on duration

We can set up a data set with one observation == one duration/spell.

- **two outcomes:**
  - **duration:** how long governments stay in power
  - **censoring:** all cabinets will end, but we don't always observe it (censor == 1).
- covariates: don't change during the spell

```
##   duration censor12
## 1        7        1
## 2       27        1
## 3        6        1
## 4       49        0
## 5        7        1
## 6        3        1
```

Time-varying covariates: focus on time

If we want to include time-varying predictors, we need to make a new observation for each change in x (and not only in y)

- we "slice" up the duration for each unit
- **four dependent variables** to account for the nesting:
  - **start** and **stop** times (i.e. duration/counter) $\rightarrow$ focus of duration models
  - occurrence of **event** (i.e. censoring) $\rightarrow$ focus of BTSCS (e.g. logits)
  - id for each **spell** that is now "sliced up"

Time-series cross-section/panel

Same as panel data

- **a fixed period for all units** i.e. all units are observed at fixed time intervals (day, week year. . .)
- **event(s)** are reported

Outcome leveraged

Outcome leveraged

We have potentially two outcomes

```
##   duration censor12
## 1        7        1
## 2       27        1
## 3        6        1
## 4       49        0
## 5        7        1
## 6        3        1
```

Focus on :

- **event** and control for duration: BTSCS approach
- **duration** (spell) punctuated by events: duration models
    - survival
    - failure

Event: Binary Time-Series-Cross-Section (BTSCS)

Event: Binary Time-Series-Cross-Section (BTSCS)

With panel data where the event is indicated, we may simply. . .

- regress the **binary event** on predictors of choice (logit, probit, log-log)
- control for the duration:
  - fixed effects for each duration (problem if we have low ratio event/no event)
  - “splines” (moving averages; hard to know what goes on)
  - cubic term
  - linear (!)

$\Rightarrow$ *ignore/treat as noise the censoring and the time between events*

Both: Duration models

Different focus in outcomes

Duration models draw information from both duration and event:

- explicit assumption that duration is partially unobserved.
- censored observations contribute with information about duration, but not event

Functional forms (duration)

Functional forms (duration)

Duration models assume a baseline probability that an event will occur that vary over time

- **accelerated failure time** (AFTs)
  - survival function
  - $S(t) = 1 - F(t) = Pr(T > t)$
  - probability that observation has lasted until now
- **proportional hazard**
  - hazard function
  - $h(t) = \frac{f(t)}{S(t)} = \frac{f(t)}{1-F(t)}$
  - probability that the observation will experience the event, given that it has lasted until now

$\Rightarrow$ *Same regression coefficients, but reverted (+/-)*

Proportional hazards

Proportional hazards

The outcome is hazard rates $h_i(t)$

- the proportion of observations that experience event in a given period
- among those that entered the period without having experienced the event

Hazards are proportional to the β

$$h_i(t) = h_o(t) exp(x_i^T \beta)$$

- $h_o(t)$ **baseline hazard varies over time:**
  - constant
  - increasing/decreasing
  - empirically determined
- $exp(x_i^T \beta)$ **single set of slope coefficients**
  - proportional to (i.e. multiplicative) the baseline hazard
  - $exp(\beta)$ reports the marginal change
    - positive β: increase in hazard/probability of event in period (decrease in duration)
    - negative β: decrase in hazard/probability of event in period (increase in duratoin)

$\Rightarrow$ *Assumption that coefficients are constant across time*

Two types of models

There are two estimation strategies/models

- **parametric models**
  - duration is continuous
  - parameters determine the shape of the baseline hazard
  - e.g. Weibull, exponential ...
- **semi-parametric model**
  - duration is ordinal
  - Cox proportional hazard

Cox proportional hazard models

Cox proportional hazard models

- order data according to event date
  - main asset: agnostic about functional form
- "partial likelihood":
  - within first date (period):
    - compare events to no-events
    - remove observations with events from data
  - within second date (period):
    - rinse and repeat
  - main weakness: **ties**
    - observations with the same event time
    - options for ordering/simulation (Efron, Breslow, exact)

Two assumptions to test in duration models

- proportional hazard: are coefficients the same over time
  - parametric and semi-parametric models
- functional form (parametric models)
- ties (Cox proportional hazard)

Dependencies between observations

Dependencies between observations

Observations are often not independent from each other in these models

- **risk set** a duration for a natural unit is sliced up
- **repeated events**
- **Time dependency** (i.e. dates)
- **different populations** split-population data

Hierarchical data structures

Different vocabularies, same thing

- strata / fixed-effects (+ clustering of errors)
- frailty / random effects
- split population / zero-inflated models

Split population model

We model two outcomes in two equations

- **the “at-risk” group** the probability of an event occurring at any given time for those who have **not** yet experienced the event
- **the “affected” group** the probability of an event occurring at any given time for those who have **already** experienced the event.