European Doctoral School of Demography 2018-19

EDSD 220 - Statistical Demography

Discrete Time Survival Models

GIANCARLO CAMARDA
Institut National d'Etudes Démographiques

December 2018
University of Southern Denmark, Odense

Discrete Time: concept

- Time is continuous but events are registered only for discrete time units, i.e. interval censoring
- Examples:
  - Age at death (in completed years/months/etc.)
  - Age at childbearing (in completed years/months/etc.)
- Failures can actually occur at discrete time units
- Examples:
  - Duration of unemployment, usually in full months
  - Number terms until graduation from university
  - Time to pregnancy if measured in menstrual cycles
- In both cases, we have a considerable number of **tied observations**

Defining Discrete Hazard

- If a random variable T can take only integer values $t = 1, 2, \ldots$, we have the following survival functions:

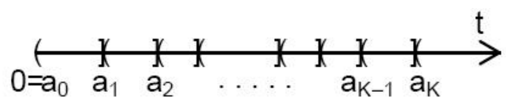
$$P(T = t) = f_t$$
$$P(T > t) = S_t$$
$$P(T = t \mid T \geq t) = \frac{P(T=t)}{P(T \geq t)} = h_t \qquad \text{discrete hazard}$$

- Note that $T \geq t \Rightarrow T > t - 1$, then:

$$h_t = P(T = t \mid T \geq t) = P(T = t \mid T > t - 1) = \frac{f_t}{S_{t-1}}$$

Continuous vs. Discrete

- In continuous time: $\tilde{T} \sim \tilde{f}(t), \tilde{S}(t), \tilde{h}(t)$
- But only observed at discrete time points, $a_1, a_2, \ldots, a_K$, i.e. data are interval censored.
- If individual is already dead at a_k, then $a_{k-1} < \tilde{T} \leq a_k$



- Discrete-time random variable:

$$T \sim f_t = P(T = t) = P(a_{t-1} < \tilde{T} \leq a_t), \qquad t = 1, \ldots, K$$

- $T = t$ stands for the event in the t^{th} interval
- We now construct a model for T based on $\tilde{f}$, $\tilde{S}$ and $\tilde{h}$

Regression Models/1

- Discrete-time hazard: $h_t = \frac{f_t}{S_{t-1}}$
- We can include covariates $\boldsymbol{x} = (x_1, \dots, x_m)'$ by the linear predictor $\boldsymbol{x}'\beta$
- Assuming $\tilde{T}$ follows continuous PH model:

$$\tilde{h}(t, x) = \tilde{h}_0(t)\, e^{\boldsymbol{x}'\beta}$$

- What does this imply for the discrete random variable T?

$$\begin{aligned} h_t &= \frac{P(T=t)}{P(T \geq t)} = \frac{P(a_{t-1} < \tilde{T} \leq a_t)}{P(\tilde{T} > a_{t-1})} \\ &= \frac{P(\tilde{T} > a_{t-1}) - P(\tilde{T} \geq a_t)}{P(\tilde{T} > a_{t-1})} = \frac{\tilde{S}(a_{t-1}) - \tilde{S}(a_t)}{\tilde{S}(a_{t-1})} = 1 - \frac{\tilde{S}(a_t)}{\tilde{S}(a_{t-1})} \end{aligned}$$

Regression Models/2

- For a continuous PH: $\tilde{S}(u) = \exp\left[-\tilde{H}_0(u)\, e^{\boldsymbol{x}'\beta}\right]$
- For a discrete PH:

$$\begin{aligned} h_t(\boldsymbol{x}) &= 1 - \frac{\tilde{S}(a_t, \boldsymbol{x})}{\tilde{S}(a_{t-1}, \boldsymbol{x})} = 1 - \frac{\exp\left[-\tilde{H}_0(a_t)\, e^{\boldsymbol{x}'\beta}\right]}{\exp\left[-\tilde{H}_0(a_{t-1})\, e^{\boldsymbol{x}'\beta}\right]} \\ &= 1 - \exp\left\{-e^{\boldsymbol{x}'\beta}\left[\tilde{H}_0(a_t) - \tilde{H}_0(a_{t-1})\right]\right\} \\ &= 1 - \exp\left\{-e^{\boldsymbol{x}'\beta} \int_{a_{t-1}}^{a_t} \tilde{h}_0(u)du\right\} \end{aligned}$$

- Let define: $\beta_{0t} = \ln \int_{a_{t-1}}^{a_t} \tilde{h}_0(u)du$

$$\Rightarrow \quad h_t(\boldsymbol{x}) = 1 - \exp\left\{-e^{\beta_{0t} + \boldsymbol{x}'\beta}\right\}$$

Setting Up the Likelihood/1

- n independent individuals
- An individual i either dies in one interval t_i or is censored, i.e. we know that $T_i > t_i$ ($i = 1, \dots, n$)
- Data $= (w_i = 0, y_i, \delta_i, \boldsymbol{x}_i)$ (for the moment no truncation)
- Usual likelihood:

$$L = \prod_{i=1}^{n} [P(T_i = y_i)]^{\delta_i} [P(T_i > y_i)]^{1-\delta_i}$$

Setting Up the Likelihood/2

- We can re-write some probabilities:
  - The probability to survive up to t, i.e. the probability to survive each interval up to t:

$$P(T > t) = \prod_{j=1}^{t} (1 - h_j)$$

  - The probability to die exactly at t, i.e. the probability to survive each interval up to $t-1$ and then die in t:

$$P(T = t) = h_t \prod_{j=1}^{t-1} (1 - h_j)$$

- And define the following abbreviation:

$$h_{it} = P(T_i = t \,|\, T_i \geq t)$$

Setting Up the Likelihood/3

- We write the log-likelihood:

$$\ln L = \sum_{i=1}^{n} \delta_i \ln \frac{h_{iy_i}}{1-h_{iy_i}} + \sum_{i=1}^{n}\sum_{j=1}^{y_i} \ln(1-h_{ij})$$

- One more step: for each individual i define a series of dummy variables d_{it}:

$$d_{it} = \begin{cases} 1 & \text{if individual dies in interval } t \\ 0 & \text{otherwise} \end{cases}$$

- The log-likelihood can be re-written:

$$\ln L = \sum_{i=1}^{n}\sum_{j=1}^{y_i} d_{ij} \ln \frac{h_{ij}}{1-h_{ij}} + \sum_{i=1}^{n}\sum_{j=1}^{y_i} \ln(1-h_{ij})$$

Bernoulli Likelihood

- D is a binary random variable (RV)
- With n independent observations $(D_1 = d_1, \dots, D_n = d_n)$

$$L = \prod_{i=1}^{n} p^{d_i}(1-p)^{1-d_i} \Rightarrow \ell = \sum_{i=1}^{n}[d_i \ln p + (1-d_i)\ln(1-p)]$$

$$= \sum_{i=1}^{n}\left[d_i \ln \frac{p}{1-p} + \ln(1-p)\right]$$

- Do you the similarities?

$$\ln L = \sum_{i=1}^{n}\sum_{j=1}^{y_i}\left\{d_{ij} \ln \frac{h_{ij}}{1-h_{ij}} + \ln(1-h_{ij})\right\}$$

- For each individual i a series of y_i observations contribute to the likelihood **as if** they were a Bernoulli RV!

Linking Covariates

- We saw that PH discrete-time model

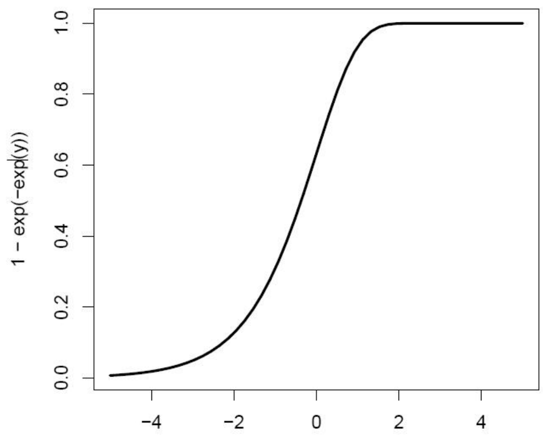
$$h_{ij} = h_j(\boldsymbol{x}_i) = P(T_i = j \mid T_i \geq j,\, \boldsymbol{x}_i) = 1 - \exp\left\{-e^{\beta_{0t} + \boldsymbol{x}'\boldsymbol{\beta}}\right\}$$

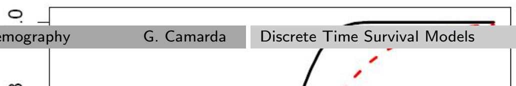
- This is also called complementary log-log link
- As alternative: logistic transformation:

$$h_{ij} = h_j(\boldsymbol{x}_i) = \frac{e^{\beta_{0j} + \boldsymbol{x}_i'\boldsymbol{\beta}}}{1 + e^{\beta_{0j} + \boldsymbol{x}_i'\boldsymbol{\beta}}}$$

- Inserting into log-likelihood $\Rightarrow$ Logistic regression model!

The Link Function





Summary

- Each individual contributes y_i new observations, the "responses" are the d_{ij}
- Several observations from one individual: what about independence? The log-likelihood is correct, we only exploit the formal identity with the Bernoulli likelihood!
- The baseline hazard is transferred into new parameters β_{0t}, $t = 1, \ldots, K$, and can be estimated
- Time-varying covariates can be easily incorporated
- Problems to anticipate:
  - Sample size!
  - Time-intervals with no events
  - Unstable likelihood because of many additional parameters

Including Left Truncation

- Data: $(w_i, y_i, \delta_i, x_i)$
- Likelihood:

$$L = \prod_{i=1}^{n} \frac{[P(T_i = y_i)]^{\delta_i}[P(T_i > y_i)]^{1-\delta_i}}{P(T_i > w_i)}$$

$$\Rightarrow \quad \ln L = \sum_{i=1}^{n} \delta_i \ln \frac{h_{iy_i}}{1 - h_{iy_i}} + \sum_{i=1}^{n}\sum_{j=1}^{y_i} \ln(1 - h_{ij}) - \sum_{i=1}^{n}\sum_{j=1}^{w_i} \ln(1 - h_{ij})$$

- If entry time $w_i > 0$, then $d_{ij} = 0$ for $j \leq w_i$

$$\Rightarrow \quad \ln L = \sum_{i=1}^{n} \sum_{j=w_i+1}^{y_i} \left[d_{ij} \ln \frac{h_{ij}}{(1 - h_{ij})} + \ln(1 - h_{ij}) \right]$$

- Each individual contributes $y_i - w_i$ Bernoulli trials.

Simulating from a PH Model

- PH model:

$$h(t) = h_0(t) e^{\mathbf{x}'\boldsymbol{\beta}}$$

- Assume we want to simulate random durations that follow this model, where $h_0(t)$ is the hazard of some parametric distribution
- Basic principle: $p \sim \text{Unif}[0,1] \Rightarrow F^{-1}(p) = t$ is a random number from wanted distribution

$$F(t) = 1 - S(t) = 1 - [S_0(t)]^{\exp\{\mathbf{x}'\boldsymbol{\beta}\}} = 1 - [S_0(t)]^r$$

Then

$$F_0(t) = \tilde{p} = 1 - (1 - p)^{\exp\{-\mathbf{x}'\boldsymbol{\beta}\}}$$

- Therefore $p \sim \text{Unif}[0,1] \rightarrow$ get $\tilde{p} \rightarrow$ invert $F_0(t)$