Prevalent cohort study

Niels Keiding
Department of Biostatistics
University of Copenhagen

Delayed entry:
some individuals observed only from some entry time

X_i iid density f, survival function S hazard α = f/S

V_i entry time U_i time of censoring X_i time of death

Observe $V_i\ , X_i \wedge U_i\ , \delta_i = \mathrm{I}\{X_i \le U_i\}$
in conditional distribution given $X_i > V_i$.

Likelihood $$\prod_{i=1}^{n} \frac{\alpha(X_i)^{\delta_i}\ S(X_i \wedge U_i)}{S(V_i)}$$

Nonparametric maximum likelihood estimator (Kaplan & Meier)

$$\hat{S}(x) = \prod_{\substack{X_i \le x \\ \delta_i = 1}} \left(1 - \frac{1}{Y(X_i)}\right)$$

$Y(x) = \#\{j \,|\, V_j < x \le X_j \wedge U_j\}$ = number at risk at x

Random delayed entry

Basic property that makes delayed entry work:

If V and X are independent then for $x>v$

hazard of X given $V = v$ and $X>V =$

$$\frac{P\{X = x | V = v,\ X > V\}}{P\{X \geq x | V = v,\ X > V\}} = \frac{P\{X = x,\ V = v\}}{P\{X \geq x,\ V = v\}}$$

$$= \frac{P\{X = x\}}{P\{X \geq x\}} = \text{hazard of } X$$

Random delayed entry: Covariates

Assume $V \perp X \mid Z$

V and X conditionally independent given Z

Then

$$\frac{P\{X = x \mid V = v,\ Z = z, X > V\}}{P\{X \geq x \mid V = v,\ Z = z, X > V\}} = \frac{P\{X = x \mid Z = z\}}{P\{X \geq x \mid Z = z\}}$$

so the *conditional hazard given* Z is estimable using the delayed entry approach under the weaker assumption that entry time V and survival time X are just *conditionally* independent given Z.

Prevalent cohort study

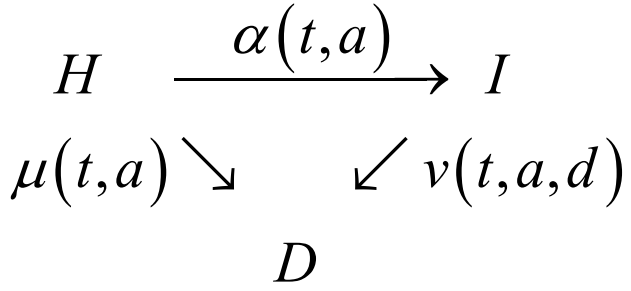
A cross-sectional sample of diseased is followed up further – usually for some fixed calendar time.

Is it correct to analyze survival of patients in a prevalent cohort using Kaplan-Meier with left truncation?

In which sense do we have ‘independent delayed entry’?

Age-status-duration distribution in cross-sectional sample

Illness-death process



calendar time t, age a, duration d

Recruited from Poisson process $\beta(t)$

In cross-sectional sample at time t, the joint distribution of current age Z, current duration Z-Y and status is

$C(t)\beta(t-z)e^{-\int_0^z(\mu(t+u-z,u)+\alpha(t+u-z,u))du}$, healthy individual

$C(t)\beta(t-z)e^{-\int_0^y(\mu(t+u-z,u)+\alpha(t+u-z,u))du}\alpha(t+y-z,y)\ e^{-\int_y^z v(t+u-z,u,u-y)du}$, diseased individual

X age at death $\quad$ Y age at onset $\quad$ T time at death

V age at entry $\quad$ T_V time at entry $= T-\left(X-V\right)$

Show for $t > t_v$, $x>y$ $\quad$ hazard $\left(X \middle| Y=y, T=t, T_V=t_v, X>V\right)$

$=$ hazard $\left(X \middle| Y=y, T=t\right)$. $\quad$ Then this is directly estimable.

$$\frac{P\left(X=x \middle| Y=y, T=t, T_V=t_v, X>V\right)}{P\left(X\geq x \middle| Y=y, T=t, T_V=t_v, X>V\right)}$$

$$= \frac{\beta\left(t-x\right)e^{-\int_0^y\left\{\mu\left(t-x+s,s\right)+\alpha\left(t-x+s,s\right)\right\}ds}\ \alpha\left(t-x+y,y\right)e^{-\int_y^x \nu\left(t-x+s,s,s-y\right)ds}\ \nu\left(t,x,x-y\right)}{\beta\left(t-x\right)e^{-\int_0^y\left\{\mu\left(t-x+s,s\right)+\alpha\left(t-x+s,s\right)\right\}ds}\ \alpha\left(t-x+y,y\right)e^{-\int_y^x \nu\left(t-x+s,s,s-y\right)ds}}$$

$$= \nu\left(t,x,x-y\right)$$

If the death intensity depends on both time t, age x, duration x-y, it has to be modelled like that.

Prevalent cohort study

Standard approach: Survival analysis with delayed entry (left truncation); risk set only includes patients *from sampling onwards*

Covariates may be included e.g. through Cox regression model

$$\lambda_i(t) = \lambda_0(t) e^{\beta \cdot Z_i(t)}$$

Alternative 1: length-biased distribution

a) patients with long disease durations are overrepresented in the prevalent cohort

b) is it not a waste to throw out the known survival time from onset to sampling?

Length-biased sampling

Basic idea easiest if

$$v(t,a,d) = v(d)$$

(lethality depends only on duration)

Then: the conditional distribution of duration D given onset age Y and current age Z has density

$$e^{-\int_{z-y}^{d} v(u)du}\, v(d)$$

$\Rightarrow$ joint distribution of *Y, Z, D*

$$c\beta e^{-\int_0^y \{\mu(u)+\alpha(u)\}du}\, \alpha(y)\, \underbrace{e^{-\int_y^z v(u-y)du} e^{-\int_{z-y}^{d} v(u)du} v(d)}_{v(d)e^{-\int_0^d v(u)du}}$$

$$\text{for } 0 < \underbrace{y < z < y + d}_{\text{interval for } z \text{ of length } d}$$

integrating out z and then y $\Rightarrow$

$$\text{const.} \cdot d\ v(d) e^{-\int_0^d v(u)du}$$

density of the length biased distribution

Estimate v from length-biased distribution (counting everybody from disease onset): under uncensored observation the NPMLE of the distribution function

$$H(d) = 1 - e^{-\int_0^d v(u)du}$$

is the Cox-Vardi estimator

$$\hat{H}(d) = \sum_i \frac{I\{D_i \le d\}}{D_i} \Big/ \sum_i \frac{1}{D_i}$$

generalized to right censoring by Vardi (1989).

Alternative 2: forward recurrence time

Sometimes we do no not know age at onset, so can we use the marginal distribution of time from sampling to death?

Under same assumptions

$$v(t,a,d) = v(d)$$

the *density* of the forward recurrence time distribution equals

$$\text{const. } e^{-\int_0^f v(u)du} = \text{const.}\{1 - H(f)\}$$

$F = D - (Z - Y)$ = forward recurrence time

NPMLE (uncensored): Grenander (1956)

generalization to right censoring by Denby & Vardi (1986)

alternative algorithm by Huang & Zhang (1994)

Remember: if duration D has

density $h(d)$ $\quad (= v(d)e^{-\int_0^f v(u)du})$

distribution function $H(d)$

then the density of

length-biased distribution

is $\quad$ const. $d\ h(d)$

forward recurrence time distribution

is $\quad$ const. $\{1 - H(f)\}$

Not clear how to incorporate age a, **time** t and other covariates into these two models.

Therefore: so far little practical applicability in typical prevalent cohort situations.

Recapitulation

If density $h(d)$, distribution function $H(d)$

Then

length-biased density	const. $d\ h(d)$
forward recurrence time density	const. $\{1 - H(d)\}$

Oakes & Dasu (1990) Biometrika

forward recurrence time hazard

$$\frac{1 - H(d)}{\int_{\alpha}^{\infty} \{1 - H(u)\}\, du} = \frac{1}{e(d)}$$

$e(x)$ = mean residual life function of D

so

prop. Hazards for forward recurrence time

$\Leftrightarrow$ prop. Mean residual life for D

Example: prevalent cohort of insulin-dependent diabetics in Denmark

All insulin-dependent diabetics (1499) in county of Fyn (population 450,000) on 1 July 1973 were identified by going through all prescriptions at the National Health Service over a five month period.

Mortality depends on age, calendar time, duration, cf. decent analysis by Andersen et al., Biometrics 1985.

Here for illustration: assume dependence only on duration.