Duration data analysis - basic concepts

Rasmus Waagepetersen
Department of Mathematics
Aalborg University
Denmark

September 6, 2024

Course topics (tentative)

- duration data - censoring and likelihoods
- estimation of the survival function and the cumulative hazard
- semi-parametric inference - Cox's partial likelihood
- model assessment
- point process/counting process approach (review)
- parametric models
- special topics:
  - time-dependent variables
  - frailty models
  - competing risks

Estimation of probability of loss given default

Risk management in banks: probability of default and probability of loss given default (default=nedskrivning eller tab).

For each customer bank records monthly default/loss status (D,L) until first loss or customer leaves bank (Q, with no loss) or date of recording.

Examples of data sets for various customers: $\neg D,\neg D,\neg D,D,D,D$
D,L $\quad$ $\neg D,D,L$ $\quad$ $\neg D,D,D,\neg D$ $\quad$ $\neg D,D,Q$
$\neg D,\neg D,\neg D,\neg D,\neg D$

How to estimate probability of loss given default ?

First restrict attention to customers with default:
$\neg D,\neg D,\neg D,D,D,D$ $\quad$ D,L $\quad$ $\neg D,D,L$ $\quad$ $\neg D,D,D,\neg D$
$\neg D,D,Q$.

Here we observe two customers with loss given default and three customers without. Estimate 40% ?

But suppose we did not observe loss for first customer because loss did not yet occur at date of recording data ? Then estimate 40% is too small !

We did perhaps not observe customer long enough → missing data

Default sequence: we denote by a default sequence, a sequence of observations initiated by a default and ending by L, Q, $\neg D$ or by D at time of recording. E.g. the data sequence $\neg D$,D,D,$\neg D$,D,D contains two default sequences D,D,$\neg D$ and D,D.

X_L: time to loss after first default in a default sequence. I.e. for sequence D,D,L, $X_L = 2$.

Similarly define X_Q and $X_{\neg D}$ as times to quit or 'recover' (return to non-default status).

Moreover define

$$T = \min\{X_L, X_Q, X_{\neg D}\}$$

as the time to either loss, quit or recover happens

Loss is obtained if loss happens first, $X_L < X_Q$ and $X_L < X_{\neg D}$

For sequences a) D,D b) D,D,$\neg D$ and D, Q, X_L is unknown: we just know a) $X_L \geq 2$ b) $X_L \geq 3$ and c) $X_L \geq 2$

Idea: factorize into conditional probabilities

Probability of loss given default is

$$P(X_L < X_Q, X_L < X_{\neg D}) = \sum_{l=1}^{\infty} P(X_L = l, X_L < X_Q, X_L < X_{\neg D})$$

$$= \sum_{l=1}^{\infty} P(X_L = l, T \geq l) = \sum_{l=1}^{\infty} P(X_L = l | T \geq l) P(T \geq l)$$

Thus enough to estimate $P(X_L = l | T \geq l)$, $l \geq 1$, and $P(T = l | T \geq l)$ since for any $k \geq 1$

$$P(T \geq k) = \prod_{l=1}^{k-1} (1 - P(T = l | T \geq l))$$

We can estimate these quantities unbiasedly for any l !

Focus now on survival function $P(T \geq l)$ and hazard functions $P(X_L = l | T \geq l)$ and $P(T = l | T \geq l)$.

These are basic concepts in duration/survival analysis !

Example: data after default for 8 customers

‘Calendar’ time - observations after default

Custm.				Now
1	D	D	D	L
2	-	-	D	D
3	L			
4	-	-	D	¬ D
5	-	Q		
6	-	D	L	
7	-	-	-	D
8	-	-	-	L

‘Customer’ time since default

Custm.				
1	D	D	D	L
2	D	D		
3	L			
4	D	¬ D		
5	Q			
6	D	L		
7	D			
8	L			

$P(X_L = 1) = 2/8$ $P(T = 1) = 3/8$ $P(X_L = 2|T \geq 2) = 1/4$
$P(T = 2|T \geq 2) = 2/4...$

$$P(\text{Loss}) = 23/32$$

(assuming $P(X_L = l) = 0$ for $l = 5, 6, \ldots$)

Customer 2 and 7 were default at the time of recording the data.

For these customers we don't know the future - they are *censored* in duration/survival analysis terminology

For a given time point l we can remove them from the sample of customers at risk for loss if they are representative of the population of customers

Does this seem a reasonable assumption ?

“klosterforsikring”

In 1872 T.N. Thiele (Danish astronomer, statistician, actuarian) engaged in designing an annuity/insurance for unmarried women (of wealthy origin).

A woman was dependent on getting married to support her living.

Parents should be able to insure a daughter against not getting married. From certain age daughter would get a yearly amount until death or marriage.

Price of insurance: expected time to death or marriage times yearly amount.

If annuity pr. year is q and T denotes time to marriage or death, then for retirement age t_R,

$$\text{price} = qE[T - t_R | T \geq t_R] P(T \geq t_R) = q\text{mrl}(t_R) S(t_R)$$

NB: in reality future payments should be discounted to get present value of future payments (inflation)

Sometimes we define survival function as $S(t) = P(T > t)$ - distinction only matters for discrete time.

mrl: mean residual life time.

T_M, T_D: times to marriage respectively death in years.

$T = \min(T_M, T_D)$.

$$\mathbb{E}[T - t_R | T \geq t_R] = \sum_{n=0}^{\infty} P(T - t_R \geq n | T \geq t_R)$$

Assuming independence $P(T \geq t) = P(T_M \geq t)P(T_D \geq t)$.

Thiele estimated $P(T_M \geq t)$ and $P(T_D \geq t)$ for $t = 1, 2, \ldots$ using parametric models and least squares from data recorded at jomfruklostre (existing homes for unmarried women).

We will return to this data set later on in an exercise.

Practical considerations

“man...ved at gøre giftermål eller ikke gifter-mål til genstand for forsikring gør sig afhængig af den forsikredes frie vilje”

This is the reason why Thiele uses data from jomfruklostre to get valid estimates of probability that insured women do not marry - insured women might or might not be less inclined to marriage than women in general, however

“Er valget mellem gift og ugift stand end utvivlsomt altid en frivillig sag, så er der naturlige bånd på denne som på enhver frihed. Og er det end muligt for enhver at fatte og at gennemføre en cølibatsbeslutning så er der dog kræfter, mægtige kræfter, der modsætte sig”

“Jeg mener også, at det vil være nødvendigt, ikke at optage interessenter i en så fremrykket alder, at det bliver let for dem eller deres familie, at danne sig et skøn om deres individuelle sandsynlighed for at blive gift”

Time to breakdown of windturbine

Vesta A/S wants to design insurance/maintenance policies. Thus need to estimate the cost of maintaining a wind turbine.

Thus need to estimate the distribution of the time from wind turbine is installed until e.g. gear box breaks down.

The wear of a turbine depends on the load that the wind turbine is exposed to - which again depends on the weather conditions: time dependent variable. Other variables (not time dependent): type of turbine, manufacturer...

Time to death of cirrhosis

In the period 1962-1969 532 patients with the diagnosis of cirrhosis joined a randomized clinical trial for which the aim was to investigate the effect of treatment with the hormone prednison.

The patients were randomly assigned to either prednison or placebo treatments.

The survival times of the patients were observed until september 1974 so that observations were right censored for patients who were alive at this date.

Discrete or continuous time ?

In practice, data are always discrete either by construction or by rounding.

Continuous time models mathematically convenient and useful if rounding of data not too severe.

E.g. Vestas and cirrhosis data analysed using continuous time models.

Common features of duration data

1. positive
2. right skewed
3. censored (mainly right censoring) - terminal event not observed at time of recording data.
4. theory very much based on probability.
5. semi-parametric methods very important.

Due to 1. and 2. normal models usually not useful.

Ignoring 3. will introduce possibly strong bias of estimates.

5. is a concept very different from usual parametric models.

Selfstudy: various parametric alternatives to normal models (exponential, Weibull, log normal, gamma).

Hazard and survival function

Let T denote random duration time with pdf f and cdf F.

Assume T continuous random variable.

Survival function

$$S(t) = P(T > t) = 1 - F(t)$$

Hazard function

$$h(t) = f(t)/S(t)$$

$h(t)\mathrm{d}t$: probability that $T \in [t, t + \mathrm{d}t[$ given $T \geq t$.

Plots of hazard function usually more informative than plots of survival function.

Types of right censoring

Let X be duration time and C time to censoring.

We observe $T = \min(X, C)$ and $\Delta = 1[X \leq C]$ ($\Delta = 1$ means duration time observed).

Type 1 censoring: an event is only observed if it occurs prior to some fixed time t_{obs}.

If a subject enters at time t_{start} then $C = t_{\text{obs}} - t_{\text{start}}$.

Progressive type 1 censoring: different subjects may have different observation times t_{obs}.

Generalized type 1 censoring: different subjects may have different starting times t_{start}.

NB: if t_{start} not controlled by experimenter then more reasonable to consider it as a random variable T_{start} in which case also C is random.

Then we may have a case of competing risk/random censoring (see later slide).

Type 2 censoring

Type 2 censoring: experiment started for n individuals at time t_{start} and terminates when duration times observed for $0 < r < n$ individuals. Then $C = X_{(r)}$.

Progressive type 2 censoring: type 2 censoring applied with $r = r_1$. After r_1 duration times observed, $n_1 \geq r_1$ individuals (including the r_1 observed) are removed from the n individuals. Then type 2 censoring applied to the remaining $n - n_1$ individuals etc.

Competing risks/random censoring

If another event happens prior to the event of interest, X is not observed. C is the duration time until the other event.

E.g. X time to death of cirrhosis and C time to death of heart attack or C time to patient leaves the study due to migration.

In practice this type of censoring is difficult unless C independent of X.

We return to competing risks in the end of the course.

NB: some authors use the term random censoring for the case where C and X are independent !

Question: what about independence of X and C in case of type 1 and 2 censoring ?

Likelihoods for duration data

Suppose we have observations (t_i, δ_i) which are realizations of (T_i, Δ_i) and $\Delta_i = 1[X_i \leq C_i]$ and the X_i are continuous random variables with density f_{X_i}.

We assume the observations are independent so it is sufficient to derive the likelihood for one observation, say (t, δ) realization of (T, Δ).

NB: KM derivations on the lower half part of page 75 very sloppy ! Their equation (3.5.5) is OK if RHS is read as pdf.

Note if T continuous random variable then (T, Δ) has density g if $P(T \leq t, \Delta = \delta) = \int_0^t g(u, \delta)\mathrm{d}u$.

Case C random and independent of X

Assume C continuous random variable with density f_C.

$$P(T \leq t, \Delta = 0) = P(C < t, X > C) = \int_0^t \int_c^\infty f_C(c) f_X(x) \mathrm{d}x \mathrm{d}c = \int_0^t f_C(c) S_X(c) \mathrm{d}c$$

Thus $g(t,0) = f_C(t) S_X(t)$. By symmetry, $g(t,1) = f_X(t) S_C(t)$.

Thus likelihood is

$$f_X(t)^\delta S_X(t)^{1-\delta} f_C(t)^{1-\delta} S_C(t)^\delta = h_X(t)^\delta S_X(t) h_C(t)^{1-\delta} S_C(t)$$

Suppose we consider a parametric family $f_X(\cdot;\theta)$ for X but $f_C(\cdot)$ is constant as a function of θ (*non-informative censoring*). Then likelihood is equivalent to

$$h_X(t;\theta)^\delta S_X(t;\theta)$$

Case C is deterministic

Suppose C is deterministic and equal to the fixed value c. Given $\delta = 0$, $T = c$ is deterministic. Given $\delta = 1$, T is continuous. Distribution of T is non-standard: a mixture of a discrete and a continuous distribution.

$$P(T = t|\delta = 0) = 1[c = t] \quad \text{and} \quad P(\delta = 0) = P(X > c) = S_x(c)$$

Hence contribution to likelihood is $1[c = t]S_x(c) = S_X(t)$ if $(t, \delta) = (c, 0)$.

Further, for $0 \leq t \leq c$

$$P(T \leq t|\delta = 1) = \frac{P(X \leq t)}{P(X \leq c)} = \frac{F_X(t)}{F_X(c)} \quad \text{and} \quad P(\delta = 1) = F_x(c)$$

Hence $P(T \leq t, \delta = 1) = F_X(t)$ with density $f_X(t)$.

Summing up, likelihood is again $f_X(t)^\delta S_X(t)^{1-\delta} = h_X(t)^\delta S_X(t)$

Mixture of discrete and continuous distribution

T has density $g(t) = f_X(t)1[t < c] + S_X(c)1[t = c]$ wrt Lebesgue $+$ point mass at c.

This in the sense that

$$P(T \leq t) = \int_0^{\min(t,c)} f_X(u)\mathrm{d}u + S_X(c)1[c \leq t]$$

Likelihood for type 2 censored data

Exercise !

Less restrictive censoring assumption: independent censoring

Terminology confusing: independent censoring is *not* the same as random censoring with X and C independent (e.g. Fleming and Harrington page 26-27 or ABGK page 51).

Informally we have independent censoring if for any time t the survival of an individual with $T \geq t$ is representative of the survival of all individuals with $X \geq t$. In other words, the information that an individual is not censored at time t does not change the distribution of the remaining survival time.

Formally

$$P(X \in [t, t+\mathrm{d}t[| X \geq t, C \geq t) = P(X \in [t, t+\mathrm{d}t[| X \geq t) = h_X(t)\mathrm{d}t$$

This is enough for non-parametric estimation of survival function (Kaplan-Meier) and Cox's partial likelihood (later).

Independent censoring continued

Counter example: suppose patients tend to leave study if their condition deteriorates - thus remaining patients with $C \geq t$ and $X \geq t$ tend to be more healthy than an arbitrary patient with $X \geq t$.

Random independent censoring trivially implies independent censoring.

Type 2 censoring is also an example of independent censoring (exercise).

Back to Spar Nord

Let X_L, X_Q, $X_{\neg D}$ denote times to either loss, quit or not default.

What we certainly can estimate from data are probabilities

$$P(X_L = l | T \geq l) = P(X_L = l | X_L \geq l, X_Q \geq l, X_{\neg D} \geq l)$$

If events $\{X_L = l\}$, $\{X_Q \geq l, X_{\neg D} \geq l\}$ are conditionally independent given $\{X_L \geq l\}$ this is equal to

$$P(X_L | X_L \geq l)$$

.

(or if $X_L, X_Q, X_{\neg D}$ are independent)

In that case we can obtain estimate of survival function for X_L by formula

$$P(X_L \geq k) = \prod_{l=1}^{k-1} (1 - P(X_L | X_L \geq l))$$