NTNU
Det skapende universitet

Survival analysis (Levetidsanalyse)
Rosner 14.8-14.11

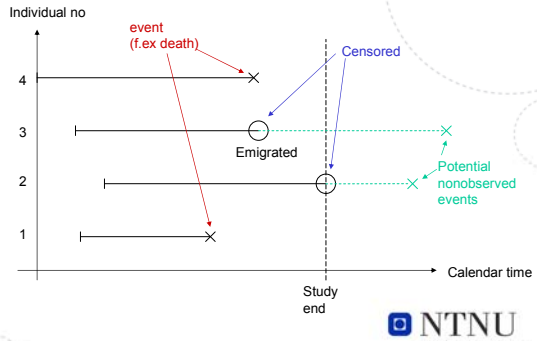
By Stian Lydersen

Lecture 21 april 2010

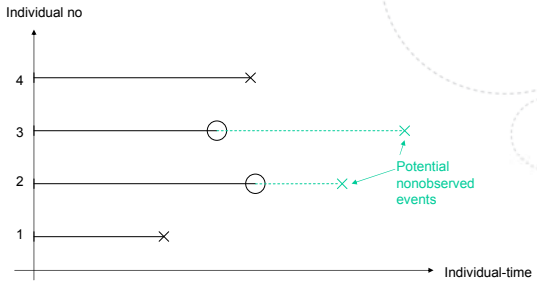
Lifetime (duration of a state)

- Examples:
  - Time to death (from birth, from diagnosis, from treatment start)
  - Time to a diagnosis
  - Time from start of treatment until declared disease-free
- Lifetimes are
  - Allways greater than 0
  - Often censored
- S(t) = Probability of surviving t

Censored data



Censored data



Lifetimie analysis – most used methods

- Actuarial calculations (approximate calculations within time intervals)
- Kaplan-Meier plot (empirical survival probability)
- Log-rank test: Non-parametric test for difference between groups
- Regression analysis:
  - Semi-parametric (Cox regression = Proportional hazard regression)
  - Parametric (Assume a certain shape of the distribution, f.ex. Weibull-distribution)

ISI-search 7 april 2010

COX DR
REGRESSION MODELS AND LIFE-TABLES
JOURNAL OF THE ROYAL STATISTICAL SOCIETY SERIES B-STATISTICAL METHODOLOGY 34 (2): 187& 1972
Times Cited: 24441

Title: NONPARAMETRIC-ESTIMATION FROM INCOMPLETE OBSERVATIONS
Author(s): KAPLAN EL, MEIER P
Source: JOURNAL OF THE AMERICAN STATISTICAL ASSOCIATION 53 (282): 457-481 1958

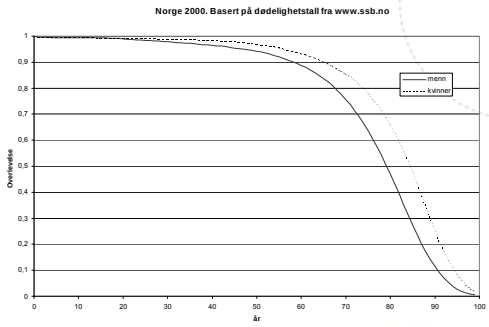
Times Cited: 33937

7

Methods for lifetime analysis assume non-informative censoring

- For each individual there exist a potential lifetime T and a censoring time C.
- We observe only min(T, C) and D, where
  - D=0 if censored
  - D=1 if event (f.ex death)
- The methods for lifetiime analysis assume T, D to be independent
- Not possible to check independence from data

8



9

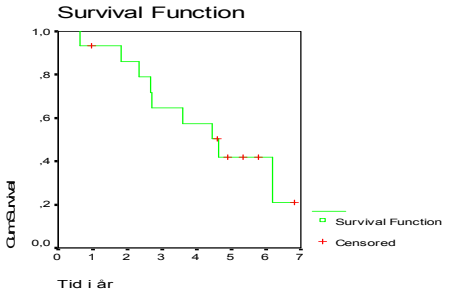
Kaplan-Meier (produkt-grense) estimatoren for overlevelsessannsynligheten

Data om overlevelse for 15 pasienter med melanom (fra Aalen, 1998)

Antall i live og under observasjon	Status (0=sensurert, 1=død)	Observasjons-tid i år	Overlevelses-sannsynlighet (Kaplan-Meier)
15	1	0,64	0,933
14	0	0,97	
13	1	1,81	0,862
12	1	2,35	0,79
11	1	2,65	0,718
10	1	2,69	0,646
9	1	3,59	0,574
8	1	4,44	0,503
7	0	4,6	
6	1	4,63	0,419
5	0	4,9	
4	0	5,32	
3	0	5,76	
2	1	6,18	0,209
1	0	6,83	

10

Kaplan-Meier plott
(empirisk overlevesessannsynlighet)



11

Comparing two groups:

H_0: No group effect
$S_1(t) = S_2(t)$ for all t

against

H_1: lifetimes tend to be longer (alternatively shorter) in Group 1
$S_1(t) \geq S_2(t)$ for all t with > for some t (alternatively ≤ og <)

Non-parametric test for complete or censored data:
Logrank test.

Approximately same result:
Test if β=0 in Cox model with x as group indicator.
(Exactly same result if no ties AND use of LR p-value in Cox model.)

12

Kleinbaum & Klein (2005) s 17
Remisjonstider for 42 leukemi-pasienter (Freireich & al, Blood, 1963)

uker	hend.	grp	uker	hend.	grp	uker	hend.	grp
1	1	2	8	1	2	16	1	1
1	1	2	8	1	2	17	0	1
2	1	2	8	1	2	17	1	2
2	1	2	8	1	2	19	0	1
3	1	2	9	0	1	20	0	1
4	1	2	10	1	1	22	1	1
4	1	2	10	0	1	22	1	2
5	1	2	11	0	1	23	1	1
5	1	2	11	1	2	23	1	2
6	1	1	11	1	2	25	0	1
6	1	1	12	1	2	32	0	1
6	1	1	12	1	2	32	0	1
6	0	1	13	1	1	34	0	1
7	1	1	15	1	2	35	0	1

Hendelse
0 = sensur
1= remisjon

Grp
1= Behandling
2=placebo

13

Lograng-test:
Eksempel fra
Kleinbaum & Klein
(2005)

Tid	hendelser		rest	
	gr1	gr2	gr1	gr2
(uke)	h1	h2	r1	r2
1	0	2	21	21
2	0	2	21	19
3	0	1	21	17
4	0	2	21	16
5	0	2	21	14
6	3	0	21	12
7	1	0	*17**	12
8	0	4	16	12
10	1	0	*15*	8
11	0	2	13	8
12	0	2	*12*	6
13	1	0	12	4
15	0	1	11	4
16	1	0	11	3
17	0	1	10	3
22	1	1	*7*	2
23	1	1	6	1
Sum	9	21		

* Sensurerte tider kommer ikke frem, *kursiv* antyder uker med sensur.

NTNU Det skapende universitet

www.ntnu.no

14

Log-rang testen

Likner på en kjikvadrat "goodness of fit" test
Del opp i korte tidsintervall
(bare ett tidspunkt med hendelse i hvert intervall)

Forventet antall hendelser i intervall nr j:
Expected = andel i risikomengden × antall hendelser totalt

$$e_{1j} = \left(\frac{r_{1j}}{r_{1j} + r_{2j}} \right) \times (h_{1j} + h_{2j}),$$

$$e_{2j} = \left(\frac{r_{2j}}{r_{1j} + r_{2j}} \right) \times (h_{1j} + h_{2j})$$

NTNU Det skapende universitet

www.ntnu.no

15

Tid	hendelser		rest		expected		
	gr1	gr2	gr1	gr2	gr1	gr2	*Gr2***
(uke)	h1	h2	r1	r2	e1	e2	Obs-Exp
1	0	2	21	21	21/42 *2	21/42 *2	1.00
2	0	2	21	19	21/40 *2	19/40 *2	1.05
3	0	1	21	17	21/38 *1	17/38 *1	0.55
4	0	2	21	16	21/37 *2	16/37 *2	1.14
5	0	2	21	14	21/35 *2	14/35 *2	1.20
6	3	0	21	12	21/33 *3	12/33 *3	-1.09
7	1	0	*17**	12	17/29 *2	12/29 *2	-0.41
8	0	4	16	12	16/28 *4	12/28 *4	2.29
10	1	0	*15*	8	15/23 *1	8/23 *1	-0.35
11	0	2	13	8	13/21 *2	8/21 *2	1.24
12	0	2	*12*	6	12/18 *2	6/18 *2	1.33
13	1	0	12	4	12/16 *1	4/16 *1	-0.25
15	0	1	11	4	11/15 *1	4/15 *1	0.73
16	1	0	11	3	11/14 *1	3/14 *1	-0.21
17	0	1	10	3	10/13 *1	3/13 *1	0.77
22	1	1	*7*	2	7/9 *2	2/9 *2	0.56
23	1	1	6	1	6/7 *2	1/7 *2	0.71
Sum	9	21			19.26	10.74	10.3

sensurerte kommer ikke frem, kursiv antyder uker med sensur

*Observert-expected for gruppe 1 = -(Observert-expected) for gruppe 2

NTNU Det skapende universitet

www.ntnu.no

16

Log-rank test

- log-rank observator: $(\Sigma_{\text{failure times}}$ obs-forv$)^2$/var (Σ obs-forv)
- var (Σ obs-forv)= $\sum \frac{r1*r2(h1+h2)*(r1+r2-h1-h2)}{(r1+r2)^2*(r1+r2-1)} = 6.3$
- Når to grupper blir sammenliknet brukes kun resultater for en gruppe; (obs-forv) for gruppe1 = - (obs-forv) for gruppe 2
- For gruppe 2: $(\Sigma_{\text{failure times}}$ obs-forv$)^2 = (10.3)^2 = 106.09$
- log-rank: chi^2= 16.8, df =1 gir p-verdi <0.0001
- Mer komplisert hvis det er mer enn 2 grupper som sammenliknes

NTNU Det skapende universitet

www.ntnu.no

17

Survival probability, survival function :
("overlevelsessansynlighet")

S(t) = P(T>t)

Hazard rate, force of mortality ("dødelighet): z(t)

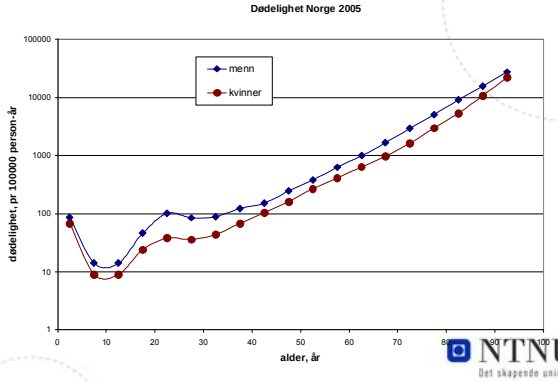
The probability that an individual of age t
dies during the next time interval of duration Δt:

$P(T \le t+\Delta t \mid T > t) \approx z(t)\,\Delta t$

NTNU Det skapende universitet

www.ntnu.no

18



NTNU Det skapende universitet

www.ntnu.no

19

Formal definition of the hazard rate:

$$z(t) = \lim_{\Delta t \to 0} \frac{P(T \le t + \Delta t \mid T > t)}{\Delta t}$$

Which is equal to

$$z(t) = \lim_{\Delta t \to 0} \frac{P((T \le t + \Delta t) \cap (T > t))}{\Delta t \quad P(T > t)} = \lim_{\Delta t \to 0} \frac{P(t < T \le t + \Delta t)}{\Delta t} \frac{1}{P(T > t)}$$

$$= \lim_{\Delta t \to 0} \frac{F(t + \Delta t) - F(t)}{\Delta t} \frac{1}{P(T > t)} = \frac{f(t)}{S(t)}$$

And also:

$$z(t) = -\frac{d}{dt}(\ln(S(t)))$$

20

Sammenheng mellom S(t) og h(t):

$$S(t) = e^{-\int_0^t h(u)du} = e^{-H(t)}$$

hvor

$$H(t) = \int_0^t h(u)du$$

kalles kumulativ hazard.

Det er lettere å plotte H(t) enn h(t).

Hazardraten h(t) = H'(t) er stigningstallet til H(t).

21

Cox Proporsjonal Hazard (PH) modell

$h(t; \underline{x}) = h_0(t) \exp(\underline{\beta}\, \underline{x})$

$h(t; x_1, \ldots, x_p) = h_0(t) \exp(\beta_1 x_1 + \ldots + \beta_p x_p)$
$= h_0(t) \exp(\beta_1 x_1) \ldots \exp(\beta_p x_p)$

ekvivalent:

$S(t; \underline{x}) = S_0(t)^{\exp(\underline{\beta}\, \underline{x})}$

hvor $h_0(t)$, $S_0(t)$ kalles "base line hazard rate", "base line overlevelsessannsynlighet"

$h_0(t)$ er ukjent og helt uspesifisert
$\beta_1, \ldots, \beta_p$ er ukjente parametre

22

An interpratation of β

- If two individuals (or groups) have respectively x_1=0 and x_1=1 and othe covariates are equal: The ratio of the hazard rates equals $\exp(\beta_1)$, independent of "age" t.
- The quantity $\exp(\beta_1)$ is called the hazard ratio (HR).
- Relevant only if the covariate is not included in an interaction term!

23

Cox, D. R. (1972):

- A method for estimation of $\beta_1, \ldots, \beta_p$
- and non-parametric estimation of $S_0(t)$ (Similar to the Kaplan-Meier estimate)
- for complete and censored data sets
- Time dependent $\beta_i(t)$ is possible.
- (Alternatively: Parametric model, for example lognormal or Weibull distribution for S(t).)

24

Partial likelihood (Cox):

$$l_p(\underline{\beta}) = \prod_{i=1}^{m} \frac{h_0(t) \exp(\underline{x}_{(i)} \underline{\beta})}{\sum_{j \in R(t_{(i)})} h_0(t) \exp(\underline{x}_j \underline{\beta})} = \prod_{i=1}^{m} \frac{\exp(\underline{x}_{(i)} \underline{\beta})}{\sum_{j \in R(t_{(i)})} \exp(\underline{x}_j \underline{\beta})}$$

where $t_{(1)} < t_{(2)} < \ldots < t_{(m)}$ aret he m distinct (uncensored) lifetimes.
$R(t_{(j)})$ is the risk set at $t_{(j)}$,
that is, individuals alive and uncensored at $t_{(j)}$
$x_{(j)}$ are the covariates for the individual with lifetime $t_{(j)}$.

25

The partial likelihood does not depend on $h_0(t)$!

Estimates for $\underline{\beta}$ as well as variances and covariances for the are obtained by treating $l_p(\underline{\beta})$ as an ordinary likelihood.

NTNU
Det skapende universitet
www.ntnu.no

26

With tied observations, a more general definition of $l_p(\underline{\beta})$ is used.

Alternatives:
- Exakt. Very time consuming.
- Breslow (1974). Quick.
- Efron (1977) Quick. Almost identical results as exact.

SPSS: Only Breslow is available

Stata: All 3 available. Breslow is default.
Use option "Efron" when tied observations.

NTNU
Det skapende universitet
www.ntnu.no

27

Checking the PH assumption: Log - log plot

We have

$$S(t,\underline{x}) = S_0(t)^{\exp(\mathbf{x}'\boldsymbol{\beta})}$$

so

$$\log(-\log(S(t,\underline{x}))) = \mathbf{x}'\boldsymbol{\beta} + \log(-\log(S_0(t)))$$

Plot for different $\underline{x}$ ought to be parallel under the PH assumption.

Difficult to judge parallelity when few observations.
More advanced methods exist.

NTNU
Det skapende universitet
www.ntnu.no

28

Software for Cox PH regression

- SPSS – has some
- MINITAB – has some
- STATA – has much
- S-plus or R- has much
- SAS - has much

NTNU
Det skapende universitet
www.ntnu.no

29

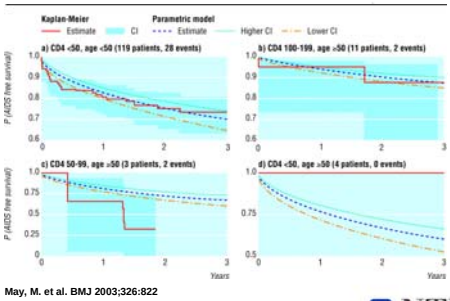
BMJ 2003;326:822
Parametric survival models may be more accurate than Kaplan-Meier estimates

EDITOR---Lundin et al use Kaplan-Meier estimates of survival probabilities in their system for survival estimation in breast cancer … They claim that researchers can obtain survival estimates based on actual data, rather than inferential estimates generated by a regression formula. However, any regression formula is based on actual data. More importantly, survival estimates from a regression model may be substantially more precise than Kaplan-Meier estimates when there are few patients in particular strata.

NTNU
Det skapende universitet
www.ntnu.no

30

Kaplan-Meier versus Weibull distribution



May, M. et al. BMJ 2003;326:822

NTNU
Det skapende universitet
BMJ
www.ntnu.no

31

References

- Hosmer, D. W., Lemeshow, D. W., May, S.: ”Applied Survival Analysis. Regression Modeling of Time to Event Data.” 2nd Ed. Wiley, 2008.
  - Comprehensive and well written book.
- Kleinbaum, D. G., and Klein, M.: ”Survival Analysis: A Self-Learning Text”. 2nd ed. Springer, 2005.
  - Introductory book, easy to read

NTNU
Det skapende universitet

www.ntnu.no