

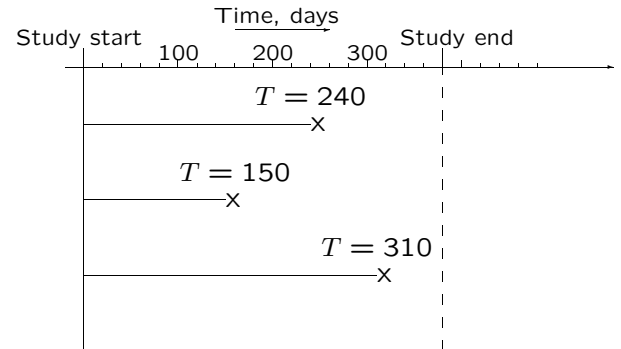
## Basic Survival Analysis

Ivan Iachine  
iachine@statdem.sdu.dk

Biostatistik - Basale Begreber  
Revision 2.11, 14.11.01

## Survival Analysis: Study of Durations Between Events

**Outcome:**  $T$  - time until an event occurs,  
"survival time" or "failure time"



*Examples:*

- Age at death
- Age at first disease diagnosis
- Waiting time to pregnancy
- Duration between treatment and death

1

The Censoring Problem in Survival Analysis:

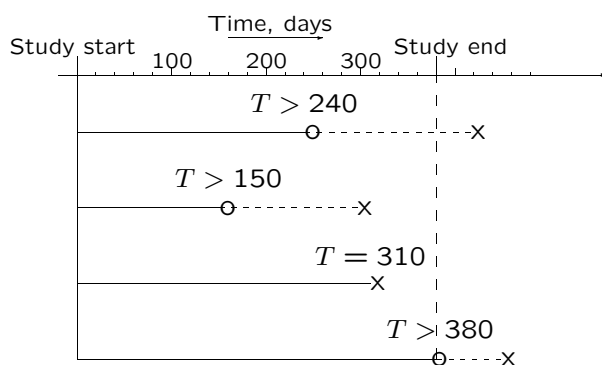
**Censoring:**

Incomplete observations of the survival time.

**Right censoring:**

Some individuals may not be observed for the full time to failure, e.g. because of:

- Loss to follow-up
- Drop out
- Termination of the study



2

*Example:*

From Kalbfleish and Prentice (1980), p.1:

"Table 1.1, from Pike (1966), gives the times from insult with the carcinogen DMBA to mortality from vaginal cancer in rats. Two groups were distinguished by a pretreatment regime." (The times are in days)

Group 1:

143, 164, 188, 188, 190, 192, 206, 209, 213, 216, 220, 227, 230, 234, 246, 265, 304, 216+, 244+

Group 2:

142, 156, 163, 198, 205, 232, 232, 233, 233, 233, 233, 239, 240, 261, 280, 280, 296, 296, 323, 204+, 344+

+Censored

3

Basic goals of survival analysis:

**1. To estimate and interpret survival characteristics:**

- Kaplan-Meier plots
- Median estimation
- Confidence intervals (CI)

**2. To compare survival in different groups:**

- Log-rank test

**3. To assess the relationship of explanatory variables to survival time:**

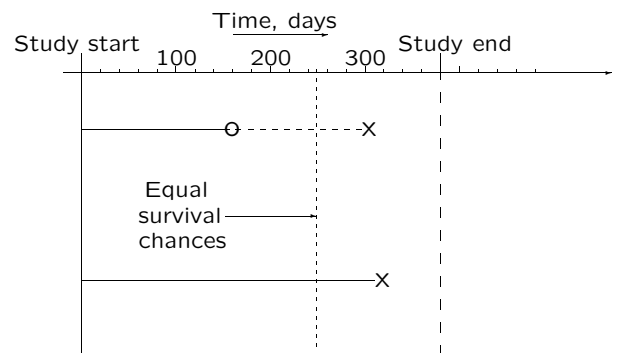
- Cox regression model

Main assumptions:

- Independent observations
- Independent censoring

Independent censoring:

Consider two random individuals both alive at time  $t$ , such that first individual is not censored and the second individual is censored at some point before  $t$ . Censoring is called *independent* if these individuals have equal chances of survival.



Examples of independent censoring:

• Simple type I censoring

All individuals are censored at the same, fixed time. Usually occurs when a population is followed during some fixed time interval.

• Progressive type I censoring

Occurs when individuals enter at different times and are followed until some fixed date. The outcome of interest is the duration between entry and event.

• Type II censoring

Occurs when following a group of  $n$  individuals and when observation ceases after  $r$ -th failure,  $r < n$ .

• Random censoring

When the time of censoring and the survival time  $T$  are independent.

Survival characteristics:

Survival function:

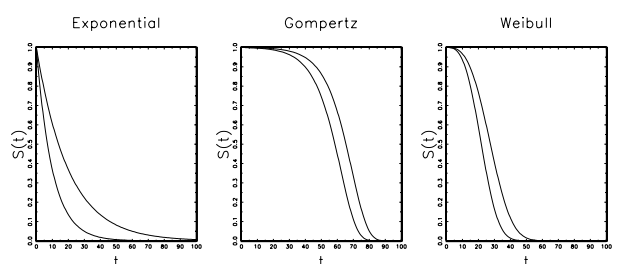
$$S(t) = P(T > t)$$

~ probability of surviving to time  $t$   
 ~ how many survive to time  $t$  in %

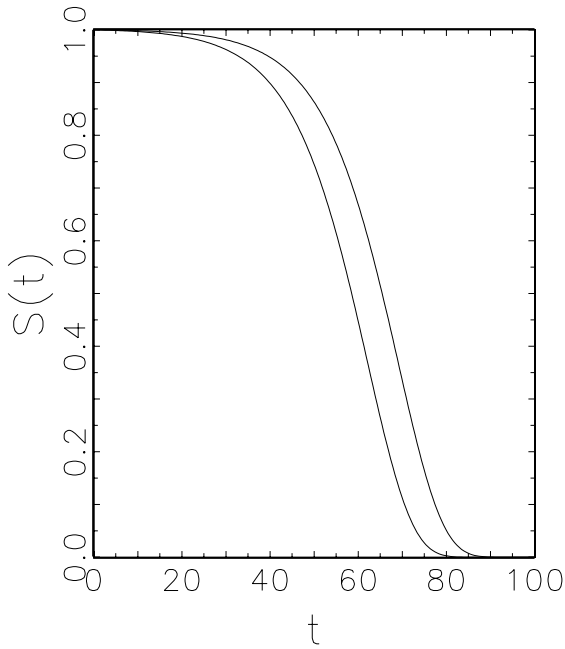
**Properties:**

- Smooth function
- $S(0) = 1$
- Decreasing function
- Cumulative survival characteristics

Examples:



### Gompertz



### Survival characteristics:

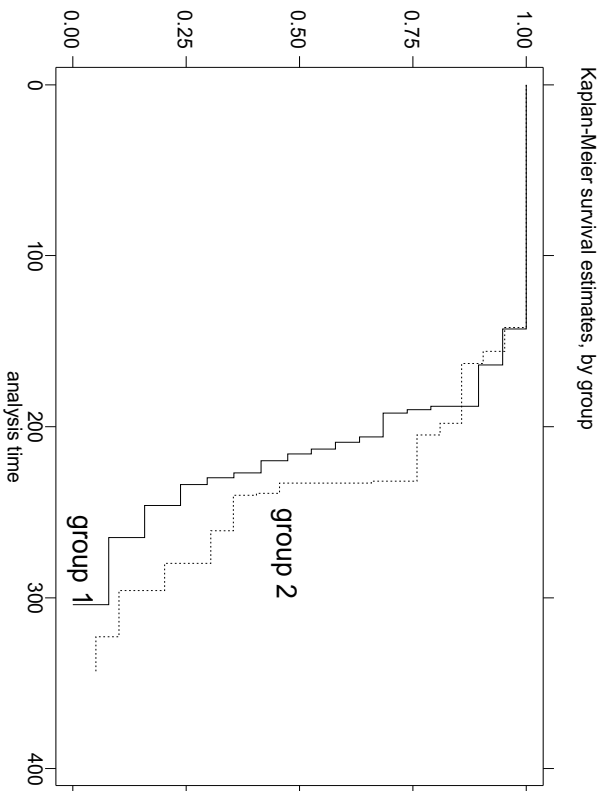
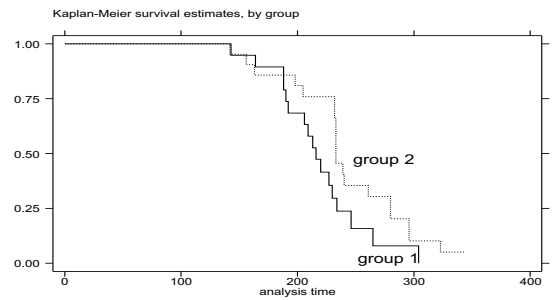
#### Kaplan-Meier (KM) estimate:

$\hat{S}(t) \sim$  estimate of  $S(t)$  from finite sample  
 $\sim$  how many in the sample survive to time  $t$  in %

#### Properties:

- Step function: Small samples=big steps
- As sample size increases:  $\hat{S}(t) \rightarrow S(t)$
- $S(0) = 1$
- Decreasing function
- Cumulative survival characteristics

#### Example:



### Population and sample properties:

Population	Finite sample
parameter	estimate of the parameter
fixed	depends on the sample
population mean $\mu$	average $\bar{x}$
survival function $S(t)$	Kaplan-Meier estimate $\hat{S}(t)$
median $t_M : S(t_M) = 1/2$	estimate of the median $\hat{t}_M : \hat{S}(\hat{t}_M) = 1/2$
regression coefficient $\beta$	estimate of the regression coefficient $\hat{\beta}$

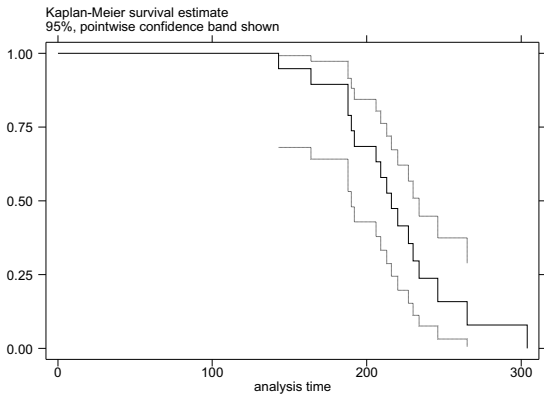
Survival characteristics:

Confidence intervals for  $S(t)$ :

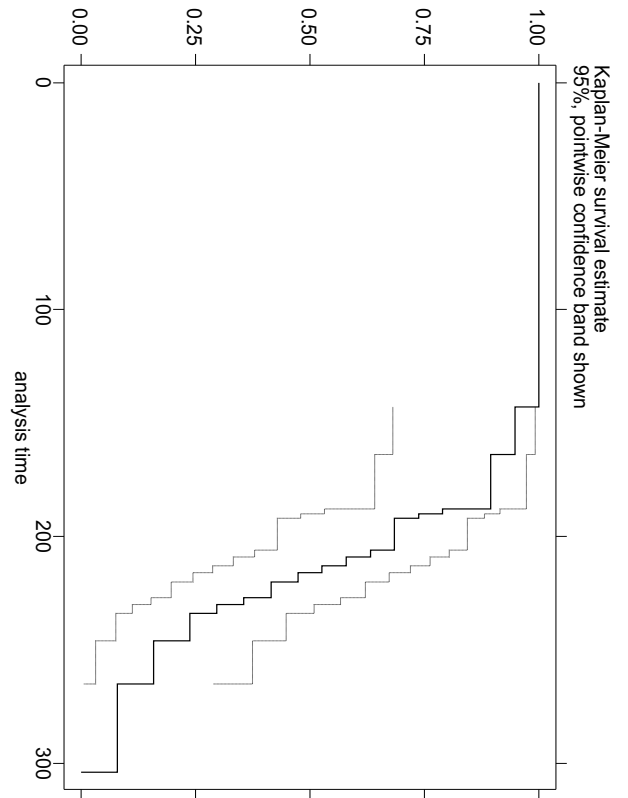
**Properties:**

- Give an idea about KM-estimate quality
- Small samples=broad confidence intervals
- *Pointwise* characteristic (for each fixed  $t$ )
- Calculated using Greenwood's formula

*Example:*



12



13

Survival characteristics:

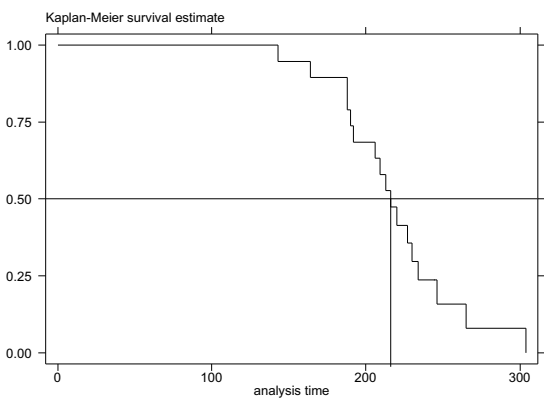
Median estimation:

**Median**,  $t_M \sim$  time when 50% are alive  
 $\sim$  satisfies  $S(t_M) = 0.5$

**Estimation via the KM-estimate:**

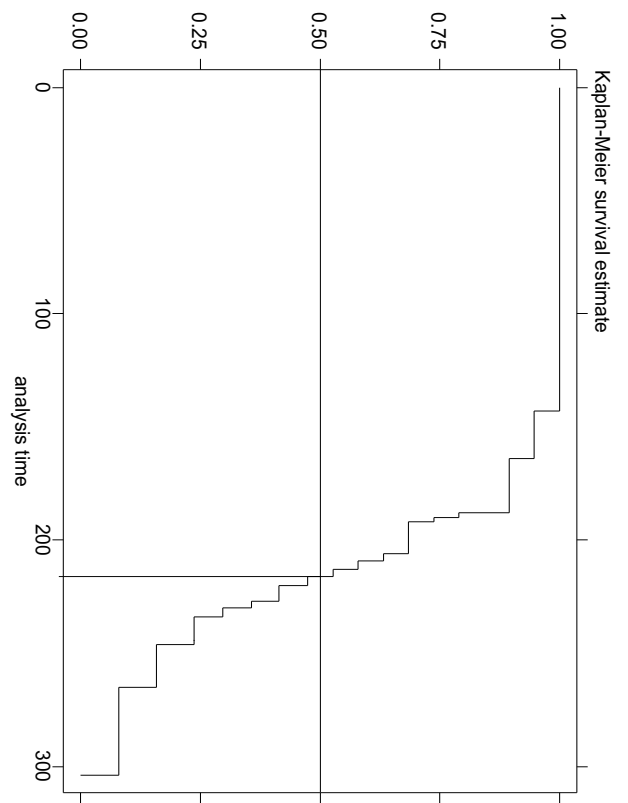
Find  $\hat{t}_M$  satisfying  $\hat{S}(\hat{t}_M) = 0.5$

*Example:*



$\hat{t}_M = 216$  days

14



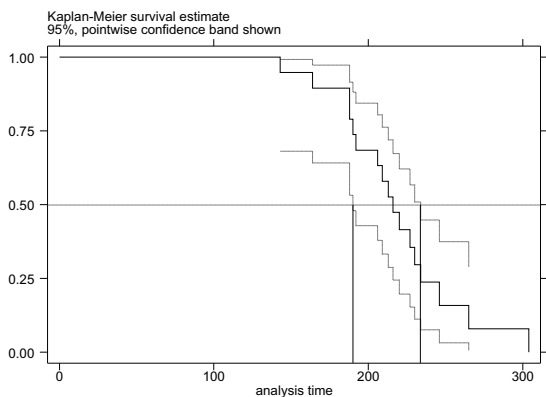
15

Survival characteristics:

Confidence intervals for the median:

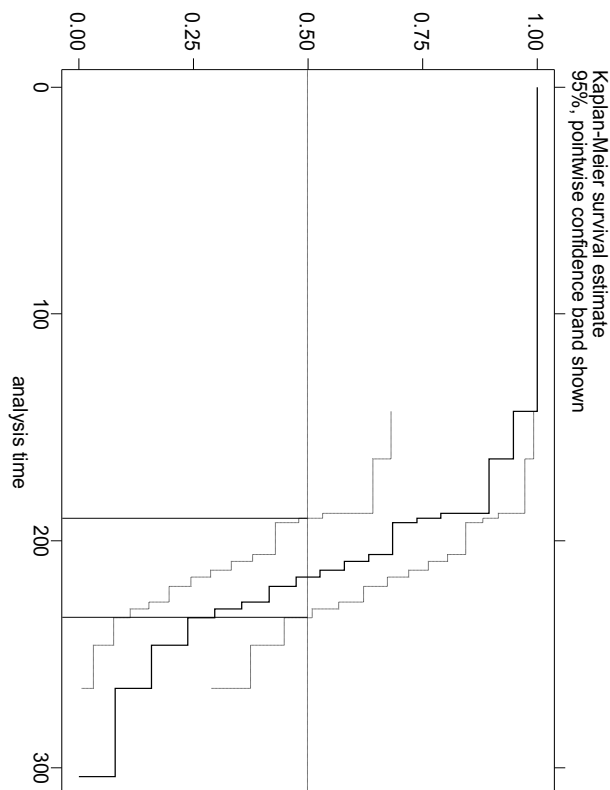
- Obtained from CI for  $S(t)$
- Small samples=broad confidence intervals
- *Note:* In general, it is wrong to test equality of two medians by comparing their CI

Example:



95%-CI for the median: [190, 234]

16



17

Comparison of survival distributions:

Two-sample logrank test:

**Group 1:** Survival function  $S_1(t)$

**Group 2:** Survival function  $S_2(t)$

**Question:**

Do the two groups have similar survival?

I.e. is  $S_1(t)$  equal to  $S_2(t)$  for all  $t$ ?

**Statistical hypothesis:**

$$H_0 : S_1 = S_2$$

$$H_A : S_1 \neq S_2$$

**Solution:**

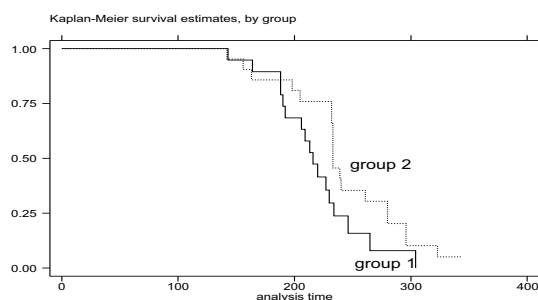
1. Choose a significance level  $\alpha$  (e.g.  $\alpha=5\%$ )
2. Compute the p-value  $P$  (e.g. using Stata)
3. If  $P \leq \alpha$  then reject the null hypothesis  $H_0$  and conclude that group 1 and 2 have different survival distributions

18

Comparison of survival distributions:

Two-sample logrank test:

Example:



Choose significance level  $\alpha = 5\%$ . Respective p-value is  $P = 0.0772 > 0.05$ , so we can not rule out that differences between the two groups are due to random sampling.

**Properties:**

- Bad performance when the two survival functions are overcrossing
- Can only be used for comparing groups defined by categorical covariates

19

**Survival characteristics:**

**Hazard function:**

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

~ slope of  $-\ln S(t)$   
 ~ how many die in % per time unit

**Interpretation:**

Consider an individual alive at time  $t$ . The chances of dying in a small interval  $[t, t + \Delta t)$  are then given by:

$$q_t = \frac{S(t) - S(t + \Delta t)}{S(t)} \approx h(t) \Delta t$$

Alternatively:

$$h(t) \approx \frac{\# \text{ of deaths in } [t, t + \Delta t)}{\# \text{ person-years at risk in } [t, t + \Delta t)}$$

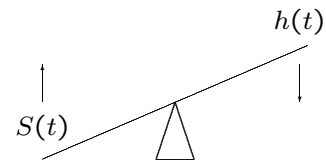
**Properties:**

- Closely related to *incidence rate*
- Not a probability, units: 1/year
- May increase or decrease or both
- *Instantaneous* survival characteristics

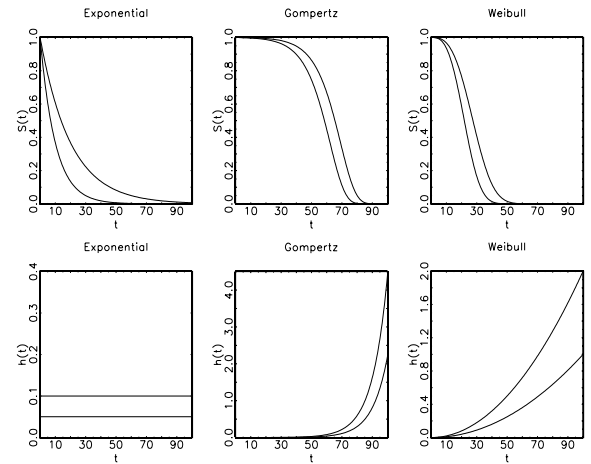
**Hazard function:**

**Properties:**

- High hazard rate = Low survival:



**Examples:**

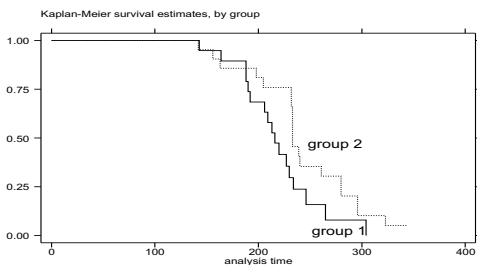


**Cox regression model:**

**Goal:**

To assess the relationship of explanatory variables (e.g. sex, age, treatment type, etc.) to survival time  $T$ .

*Example:* One covariate  $x = \begin{cases} 0, & \text{if group}=1 \\ 1, & \text{if group}=2 \end{cases}$



**One idea (Cox):**

Use a *proportional hazards* regression model:

$$h(t|x) = h_0(t) \cdot e^{\beta x}$$

where:

- $h_0(t)$  is a *baseline hazard function*
- $\beta$  is a *regression coefficient*

**Cox regression model:**

**What does it mean:**

$$h(t|x) = h_0(t) \cdot e^{\beta x}$$

**It means that:**

- In group 1, the hazard function is  $h(t|x=0) = h_0(t) \cdot e^{\beta \cdot 0} = h_0(t)$
- In group 2, the hazard function is  $h(t|x=1) = h_0(t) \cdot e^{\beta \cdot 1} = h_0(t)e^{\beta}$
- The *relative risk* for group 2 vs. group 1 is  $RR = \frac{h(t|x=1)}{h(t|x=0)} = \frac{h_0(t)e^{\beta}}{h_0(t)} = e^{\beta}$

**Interpretation of  $\beta$ 's sign:**

- $\beta > 0$  :  $RR > 1$  and  $h(t|x=1) > h(t|x=0)$
- $\beta = 0$  :  $RR = 1$  and  $h(t|x=1) = h(t|x=0)$
- $\beta < 0$  :  $RR < 1$  and  $h(t|x=1) < h(t|x=0)$

**Statistical hypothesis:**

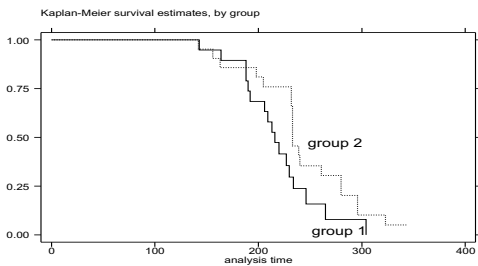
- $H_0$  :  $\beta = 0$  (no influence of covariate)
- $H_A$  :  $\beta \neq 0$  (covariate influences survival)

## Cox regression model:

Using a statistical package we can get:

- $\hat{\beta}$  - an estimate of  $\beta$
- $\hat{RR}$  - an estimate of  $RR$
- standard errors and confidence intervals
- p-values for testing hypotheses  $H_0 : \beta = 0$

Example: One covariate  $x = \begin{cases} 0, & \text{if group}=1 \\ 1, & \text{if group}=2 \end{cases}$



**Results:** (std.err. in parentheses)

- $\hat{\beta} = -0.60(0.35)$ , 95% CI [-1.28, 0.09]
- $\hat{RR} = 0.55(0.20)$ , 95% CI [0.28, 1.09]
- p-value for  $H_0 : \beta = 0$  is  $P = 0.087 > 0.05$ ,  $H_0$  cannot be rejected at 5% level

24

## Cox regression model:

Several covariates:

$$x_1, x_2, \dots, x_m$$

E.g. sex, age, treatment group, BP, BMI etc.

As before:

Use a *proportional hazards* regression model:

$$h(t|x_1, \dots, x_m) = h_0(t) \cdot e^{\beta_1 x_1 + \beta_2 x_2 + \dots + \beta_m x_m}$$

where:

- $h_0(t)$  is a *baseline hazard function*
- $\beta_1, \beta_2, \dots, \beta_m$  are *regression coefficients*

Meaning of  $\beta_j$ :

- group 1: Covariates  $x_1, \dots, x_j, \dots, x_m$
- group 2: Covariates  $x_1, \dots, x_j + 1, \dots, x_m$

The relative risk for group 2 vs. group 1 is

$$RR = e^{\beta_j}$$

Does not depend on  $x_1, \dots, x_m$ —no *interaction*

25

## Cox regression model:

Interpretation of  $\beta_j$ 's sign:

$$\beta_j > 0 : x_j \nearrow \text{ implies } h(t|x_j) \nearrow$$

$$\beta_j = 0 : x_j \text{ has no effect on } h(t|x_j)$$

$$\beta_j < 0 : x_j \nearrow \text{ implies } h(t|x_j) \searrow$$

Statistical hypothesis:

$$H_0 : \beta_j = 0 \quad (\text{no influence of } x_j \text{ on survival})$$

$$H_A : \beta_j \neq 0 \quad (x_j \text{ influences survival})$$

**Remarks:**

- no need to know  $h_0(t)$  to estimate  $\beta$
- *proportionality of hazards*

$$h(t|x) = h_0(t) \cdot e^{\beta x}$$

is an **assumption** which may or may not be true (requires *diagnostics*)
- cannot use linear regression instead of Cox because of censoring and non-symmetry of  $T$ 's distribution

26

## Cox regression model:

Example: Survival of the Danish Twins

The Danish Twin Register contains survival data on all same-sex twin pairs born in Denmark between 1870 and 1930 in which both individuals survived to age 6 (Hauge 1981, Kyvik et al. 1996). The twins were ascertained through a manual search of all birth registers kept locally by the parishes in Denmark. The zygosity diagnosis was based on a questionnaire regarding physical similarity. Only information on twin #1 in all pairs is used in this example.

**Variables:**

Survival time,  $T$ : age at death (years)  
 Covariate  $x_1$ : sex (0=males, 1=females)  
 Covariate  $x_2$ : birth cohort (0=1870, ..., 60=1930)  
 Covariate  $x_3$ : zygosity (0=MZ, 1=DZ)

**Regression model:**

$$h(t|x_1, x_2, x_3) = h_0(t) \cdot \exp(\beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3)$$

27

Cox regression model:

Example, continued:

Sample:

- 4591 male twins (1613 MZ, 2978 DZ)
- 4985 female twins (1711 MZ, 3274 DZ)

Results, regression coefficients:

		$\hat{\beta}_j$	s.e. ( $\hat{\beta}_j$ )	95% CI	p-value
sex	$x_1$	-0.36	0.025	[-0.412, -0.314]	<0.001
coh	$x_2$	-0.01	0.001	[-0.012, -0.008]	<0.001
zyg	$x_3$	0.06	0.026	[0.013, 0.115]	0.014

Results, relative risks (hazard ratios):

		$\hat{RR}_j$	s.e. ( $\hat{RR}_j$ )	95% CI
sex	$x_1$	0.70	0.017	[0.662, 0.730]
coh	$x_2$	0.99	0.001	[0.988, 0.992]
zyg	$x_3$	1.07	0.028	[1.013, 1.122]

Estimated regression model:

$$h(t|x_1, x_2, x_3) = h_0(t) \cdot \exp(-0.36x_1 - 0.01x_2 + 0.06x_3)$$

Example (TWINS, contd.):

Interpreting the estimated model I:

- $x_1$ : sex (0=males, 1=females)
- $x_2$ : byear0 - birth year (0=1870, ..., 60=1930)
- $x_3$ : zyg - zygosity (0=MZ, 1=DZ)

$$h(t|x_1, x_2, x_3) = h_0(t) \cdot \exp(-0.36x_1 - 0.01x_2 + 0.06x_3)$$

- estimated hazard for male MZ twins born in 1870 ( $x_1 = 0, x_2 = 0, x_3 = 0$ ):  

$$h(t|x_1 = 0, x_2 = 0, x_3 = 0) = h_0(t) \cdot \exp(-0.36 \cdot 0 - 0.01 \cdot 0 + 0.06 \cdot 0) = h_0(t)$$

I.e. this is a *baseline* group

- estimated hazard for male DZ twins born in 1870 ( $x_1 = 0, x_2 = 0, x_3 = 1$ ):  

$$h(t|x_1 = 0, x_2 = 0, x_3 = 1) = h_0(t) \cdot \exp(-0.36 \cdot 0 - 0.01 \cdot 0 + 0.06 \cdot 1) = h_0(t) \exp(0.06) = h_0(t) \cdot 1.06$$
- the *relative risk* for male DZ twins born in 1870 vs. the *baseline* is **1.06**

Example (TWINS, contd.):

Interpreting the estimated model II:

- $x_1$ : sex (0=males, 1=females)
- $x_2$ : byear0 - birth year (0=1870, ..., 60=1930)
- $x_3$ : zyg - zygosity (0=MZ, 1=DZ)

$$h(t|x_1, x_2, x_3) = h_0(t) \cdot \exp(-0.36x_1 - 0.01x_2 + 0.06x_3)$$

- estimated hazard for male MZ twins born in 1870 ( $x_1 = 0, x_2 = 0, x_3 = 0$ ):  

$$h(t|x_1 = 0, x_2 = 0, x_3 = 0) = h_0(t) \cdot \exp(-0.36 \cdot 0 - 0.01 \cdot 0 + 0.06 \cdot 0) = h_0(t)$$

I.e. this is a *baseline* group

- estimated hazard for male MZ twins born in 1930 ( $x_1 = 0, x_2 = 60, x_3 = 0$ ):  

$$h(t|x_1 = 0, x_2 = 60, x_3 = 0) = h_0(t) \cdot \exp(-0.36 \cdot 0 - 0.01 \cdot 60 + 0.06 \cdot 0) = h_0(t) \exp(-0.60) = h_0(t) \cdot 0.55$$
- the *relative risk* for male MZ twins born in 1930 vs. the *baseline* is **0.55**

Example (TWINS, contd.):

Interpreting the estimated model III:

- $x_1$ : sex (0=males, 1=females)
- $x_2$ : byear0 - birth year (0=1870, ..., 60=1930)
- $x_3$ : zyg - zygosity (0=MZ, 1=DZ)

$$h(t|x_1, x_2, x_3) = h_0(t) \cdot \exp(-0.36x_1 - 0.01x_2 + 0.06x_3)$$

- estimated hazard for male MZ twins born in 1870 ( $x_1 = 0, x_2 = 0, x_3 = 0$ ):  

$$h(t|x_1 = 0, x_2 = 0, x_3 = 0) = h_0(t) \cdot \exp(-0.36 \cdot 0 - 0.01 \cdot 0 + 0.06 \cdot 0) = h_0(t)$$

I.e. this is a *baseline* group

- estimated hazard for female DZ twins born in 1930 ( $x_1 = 1, x_2 = 60, x_3 = 1$ ):  

$$h(t|x_1 = 1, x_2 = 60, x_3 = 1) = h_0(t) \cdot \exp(-0.36 \cdot 1 - 0.01 \cdot 60 + 0.06 \cdot 1) = h_0(t) \exp(-0.9) = h_0(t) \cdot 0.41$$
- the *relative risk* for female DZ twins born in 1930 vs. the *baseline* is **0.41**

Advanced survival analysis:  
(Or topics not covered):

- **Left truncation:**

Selective sampling, i.e. a survival time is included in the sample if a specific condition (e.g.  $T > t_0$ ) is satisfied

- **Interval censoring:**

Information about the survival time is in the form  $t_1 < T < t_2$  (i.e. even more reduced than with right censoring)

- **Time-dependent covariates:**

Cox regression may be extended to include covariates which change with time (e.g. blood pressure  $Y(t)$ )

- **Non-proportional hazards models:**

Other types of regression models (e.g. *accelerated failure time models*) can be used to handle this situation

- **Dependent survival times:**

Bivariate survival models (e.g. *correlated frailty models*) can be used to analyze survival data on twins and relatives

32

Additional reading:

Kleinbaum D.G. (1996). *Survival Analysis: A Self-Learning Text*. Springer, New York.

Miller (1981). *Survival Analysis*. Wiley, New York.

Cox and Oakes (1984). *Analysis of Survival Data*. Chapman and Hall.

Kalbfleisch and Prentice (1980). *The Statistical Analysis of Failure Time Data*. Wiley, New York.

Lee (1980). *Statistical Methods for Survival Data Analysis*. Lifetime Learning Publications, Belmont, California.

Andersen, Borgan, Gill and Keiding (1993). *Statistical Models Based on Counting Processes*. Springer-Verlag, New York.

33

Dataset references:

Hauge, M. (1981). The Danish twin register. In Mednik, S. A., Baert, A.E., Backmann, B. P. (eds): *Prospective Longitudinal Research, An Empirical Basis for the Primary Prevention of Physiological Disorders*, London Oxford University Press pp.218-221.

Kyvik, K.O., Christensen K., Skytthe A., Harvald B. and N.V. Holm (1996). The Danish Twin Register. *Dan. Med. Bull.* 1996 Dec; 43(5): 467-470.

34