

Regression models for survival data

Assume that we have a sample of n individuals, and let $N_i(t)$ count the **observed** occurrences of the event of interest for individual i as a function of (study) time t

We have the decomposition

$$\underbrace{dN_i(t)}_{\text{observation}} = \underbrace{\lambda_i(t)dt}_{\text{signal}} + \underbrace{dM_i(t)}_{\text{noise}}$$

We will consider regression models where the intensity process λ_i for individual i depends on a vector of covariates

$$\mathbf{x}_i = (x_{i1}, x_{i2}, \dots, x_{ip})^T$$

The intensity process for individual i may be given as

$$\lambda_i(t) = \underbrace{Y_i(t)}_{\text{at risk indicator}} \cdot \underbrace{\alpha(t \mid \mathbf{x}_i)}_{\text{hazard rate}}$$

A regression model specifies how the hazard rate depends on covariates

We will consider two regression models:

- Cox's proportional hazards regression model
- Aalen's additive hazards regression model

In general the covariates in these models may depend on time

But we will mainly consider the situation with **fixed covariates** (which are measured in advance, i.e. at time zero, and remain fixed throughout the study)

Cox's proportional hazards regression model

Assume that the hazard rate for individual i takes the form

$$\alpha(t|\mathbf{x}_i) = \underbrace{\alpha_0(t)}_{\text{baseline hazard}} \underbrace{\exp(\boldsymbol{\beta}^T \mathbf{x}_i)}_{\text{hazard ratio (relative risk)}}$$

Here $\boldsymbol{\beta} = (\beta_1, \beta_2, \dots, \beta_p)^T$ is a vector of regression coefficients and the baseline hazard $\alpha_0(t)$ is the hazard for an individual with all covariates equal to zero

We make no assumptions of the form of the baseline hazard

Thus the model contains a nonparametric part (the baseline hazard) and a parametric part (the relative risk function)

We say that the model is **semiparametric**

The assumptions underlying Cox's regression model are

1) Log-linearity:

$$\log\{\alpha(t|\mathbf{x})\} = \log\{\alpha_0(t)\} + \boldsymbol{\beta}^T \mathbf{x}$$

2) Proportional hazards:

$$\frac{\alpha(t|\mathbf{x}_2)}{\alpha(t|\mathbf{x}_1)} = \exp\{\boldsymbol{\beta}^T (\mathbf{x}_2 - \mathbf{x}_1)\} \quad (\text{independent of time})$$

These assumptions should be checked when applying Cox's model, but we do not have time to look carefully at model checking in this short course

In order to understand the interpretation of the regression coefficients $\beta_1, \dots, \beta_p$, we consider two individuals, indexed 1 and 2, and assume that all components of $\mathbf{x}_1$ and $\mathbf{x}_2$ are equal, except the j -th component where $x_{2j} = x_{1j} + 1$

Then we have

$$\frac{\alpha(t | \mathbf{x}_2)}{\alpha(t | \mathbf{x}_1)} = \frac{\alpha_0(t) \exp(\boldsymbol{\beta}^T \mathbf{x}_2)}{\alpha_0(t) \exp(\boldsymbol{\beta}^T \mathbf{x}_1)} = \exp\{\boldsymbol{\beta}^T (\mathbf{x}_2 - \mathbf{x}_1)\} = e^{\beta_j}$$

Thus e^{β_j} is the hazard ratio for one unit's increase in the j -th covariate, keeping all other covariates constant

Partial likelihood and estimation of β

Ordinary ML-estimation does not work for Cox's regression model due to the nonparametric baseline

Instead we use a **partial likelihood**

We will see how this may be derived

The intensity process of N_i is given as

$$\lambda_i(t) = Y_i(t) \alpha(t | \mathbf{x}_i) = Y_i(t) \alpha_0(t) \exp(\boldsymbol{\beta}^T \mathbf{x}_i)$$

The intensity process of the aggregated counting process $N_{\cdot}(t) = \sum_{l=1}^n N_l(t)$ takes the form (assuming no joint events)

$$\lambda_{\cdot}(t) = \sum_{l=1}^n \lambda_l(t) = \sum_{l=1}^n Y_l(t) \alpha_0(t) \exp(\boldsymbol{\beta}^T \mathbf{x}_l)$$

We consider the conditional probability of observing an event for individual i at time t , given the past $\mathcal{F}_{t-}$ and given that an event is observed at time t :

$$\begin{aligned}
 \pi(i | t) &= P(dN_i(t) = 1 \mid dN_{\bullet}(t) = 1, \mathcal{F}_{t-}) \\
 &= \frac{P(dN_i(t) = 1 \mid \mathcal{F}_{t-})}{P(dN_{\bullet}(t) = 1 \mid \mathcal{F}_{t-})} \\
 &= \frac{\lambda_i(t)}{\lambda_{\bullet}(t)} \\
 &= \frac{Y_i(t) \exp(\boldsymbol{\beta}^T \mathbf{x}_i)}{\sum_{l=1}^n Y_l(t) \exp(\boldsymbol{\beta}^T \mathbf{x}_l)}
 \end{aligned}$$

We obtain the partial likelihood by multiplying together these conditional probabilities over all observed event times

Thus, if i_j is the index of the individual who experiences an event at time T_j , the partial likelihood becomes

$$\begin{aligned} L(\boldsymbol{\beta}) &= \prod_{T_j} \pi(i_j | T_j) := \prod_{T_j} \frac{Y_{i_j}(T_j) \exp(\boldsymbol{\beta}^T \mathbf{x}_{i_j})}{\sum_{l=1}^n Y_l(T_j) \exp(\boldsymbol{\beta}^T \mathbf{x}_l)} \\ &= \prod_{T_j} \frac{\exp(\boldsymbol{\beta}^T \mathbf{x}_{i_j})}{\sum_{l \in \mathcal{R}_j} \exp(\boldsymbol{\beta}^T \mathbf{x}_l)} \end{aligned}$$

where $\mathcal{R}_j = \{l | Y_l(T_j) = 1\}$ is the **risk set** at time T_j

Using results for counting processes and martingales, it may be shown that the maximum partial likelihood estimator $\hat{\boldsymbol{\beta}}$ enjoys "the usual properties" of ML-estimators

Thus $\hat{\boldsymbol{\beta}}$ is approximately multivariate normally distributed around the true value of $\boldsymbol{\beta}$ with a covariance matrix that may be estimated by $\mathbf{I}(\hat{\boldsymbol{\beta}})^{-1}$, where

$$\mathbf{I}(\boldsymbol{\beta}) = \left\{ -\frac{\partial^2}{\partial \beta_h \partial \beta_j} \log L(\boldsymbol{\beta}) \right\}$$

is the **observed information matrix**

Also Wald tests, score tests and likelihood ratio tests may be used in a similar manner as for ordinary maximum likelihood estimation

In particular, to test the null hypothesis $H_0: \beta_j = 0$, we may use the **Wald test statistic**

$$Z = \frac{\hat{\beta}_j}{\text{se}(\hat{\beta}_j)}$$

which is approximately standard normally distributed under the null hypothesis

To obtain a confidence interval for the hazard ratio e^{β_j} we transform the limits of the standard confidence interval for β_j to get the 95% confidence interval

$$\exp\{\hat{\beta}_j \pm 1.96 \text{ se}(\hat{\beta}_j)\}$$

Melanoma data

Model with sex as only covariate:

	$\hat{\beta}$	$\widehat{HR}$	$se(\hat{\beta})$	Z	P
Sex:	0.662	1.94	0.265	2.50	0.013

95% confidence interval for the hazard ratio e^{β}

$$\exp\{0.662 \pm 1.96 \cdot 0.265\}$$

which gives

$$[1.15, 3.26]$$

Model with sex and log-thickness as covariates

	$\hat{\beta}$	$\widehat{HR}$	$se(\hat{\beta})$	Z	P
Sex:	0.459	1.58	0.267	1.72	0.085
$\text{Log}_2(\text{thickness})$	0.541	1.72	0.109	4.96	$7 \cdot 10^{-7}$

We use base 2 logarithms to ease the interpretation of the estimate of the hazard ratio

R-commands that give the results on the previous slides

Read data and load survival library

```
mdata="https://www.mn.uio.no/math/english/people/aca/borgan/iwsm2026/melanoma.txt"
melanoma=read.table(mdata,header=T)
library (survival)
```

We first consider the model with sex as the only covariate:

```
fit.s = coxph(Surv(lifetime,status==1) ~ sex, data=melanoma)
summary(fit.s)
```

Note that since sex is a binary covariate (coded 1 and 2), we get the same
estimate if we treat sex as a numeric covariate or as a categorical
covariate [by using factor(sex) in the coxph-command]

We then consider the model with sex and log-thickness and as covariates:

```
fit.st = coxph(Surv(lifetime,status==1) ~ sex + log2(thickn),data=melanoma)
summary(fit.st)
```

Estimation of cumulative hazards and survival probabilities

We will estimate the cumulative baseline hazard

$$A_0(t) = \int_0^t \alpha_0(u) du$$

We take the aggregated counting process $N_{\cdot}(t) = \sum_{l=1}^n N_l(t)$ as our starting point. Its intensity process is given by

$$\lambda_{\cdot}(t) = \sum_{l=1}^n Y_l(t) \alpha_0(t) \exp(\boldsymbol{\beta}^T \mathbf{x}_l) = \{\sum_{l=1}^n Y_l(t) \exp(\boldsymbol{\beta}^T \mathbf{x}_l)\} \alpha_0(t)$$

The intensity process has a multiplicative form

So if we had known $\boldsymbol{\beta}$, we could have estimated $A_0(t)$ by a Nelson-Aalen type estimator

For a given value of $\boldsymbol{\beta}$, we may estimate $A_0(t)$ by

$$\hat{A}_0(t; \boldsymbol{\beta}) = \int_0^t \frac{dN_{\bullet}(u)}{\sum_{l=1}^n Y_l(u) \exp(\boldsymbol{\beta}^T \mathbf{x}_l)}$$

Since $\boldsymbol{\beta}$ is unknown, we replace it by $\hat{\boldsymbol{\beta}}$ to obtain the **Breslow estimator**:

$$\begin{aligned} \hat{A}_0(t) &= \int_0^t \frac{dN_{\bullet}(u)}{\sum_{l=1}^n Y_l(u) \exp(\hat{\boldsymbol{\beta}}^T \mathbf{x}_l)} \\ &= \sum_{T_j \leq t} \frac{1}{\sum_{l \in \mathcal{R}_j} \exp(\hat{\boldsymbol{\beta}}^T \mathbf{x}_l)} \end{aligned}$$

where $T_1 < T_2 < \dots$ are the event times

The cumulative hazard corresponding to an individual with a given covariate vector $\mathbf{x}_0$ is

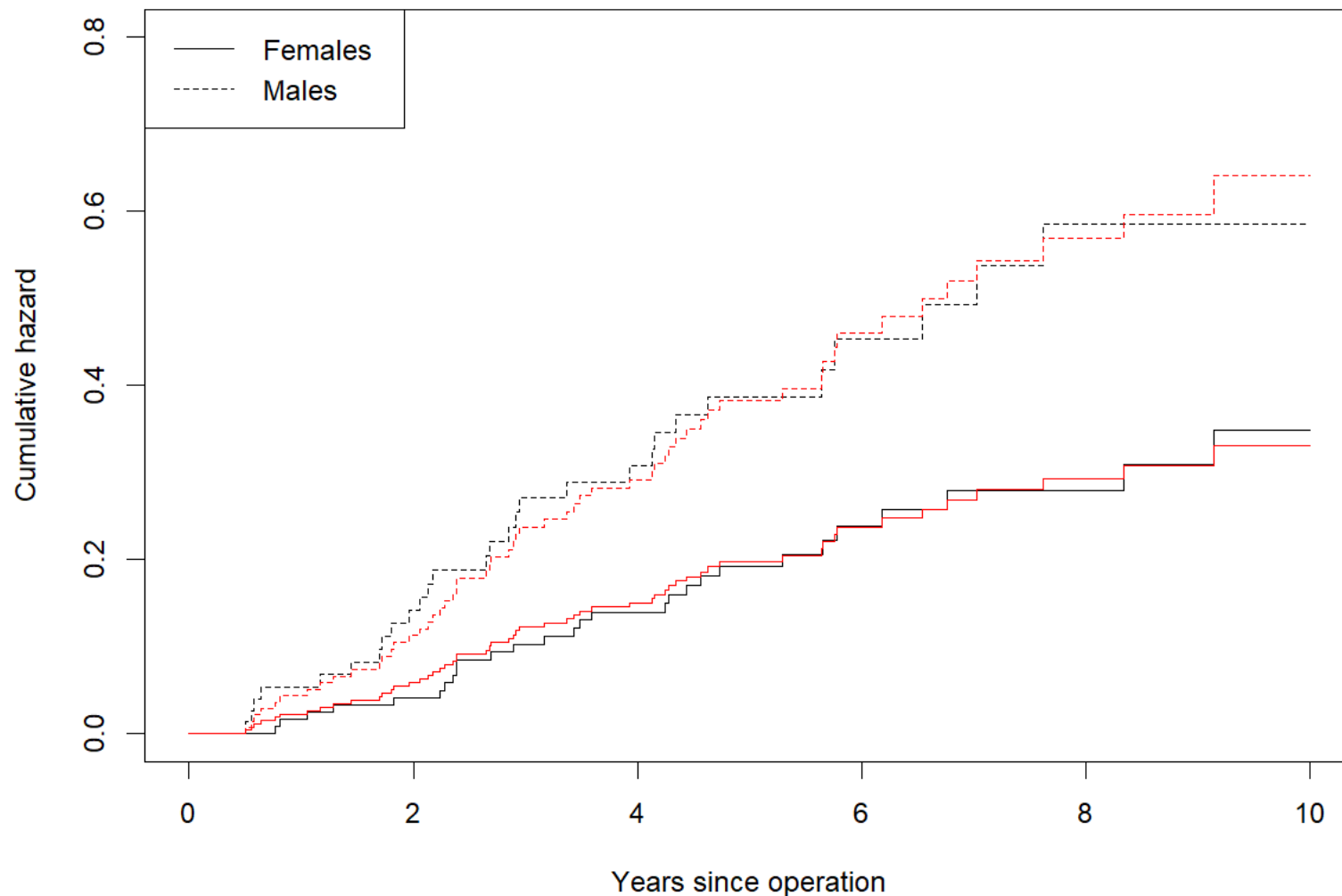
$$\begin{aligned} A(t | \mathbf{x}_0) &= \int_0^t \alpha(u | \mathbf{x}_0) du \\ &= \int_0^t \exp(\boldsymbol{\beta}^T \mathbf{x}_0) \alpha_0(u) du \\ &= \exp(\boldsymbol{\beta}^T \mathbf{x}_0) A_0(t) \end{aligned}$$

and it may be estimated by

$$\hat{A}(t | \mathbf{x}_0) = \exp(\hat{\boldsymbol{\beta}}^T \mathbf{x}_0) \hat{A}_0(t)$$

Illustration: melanoma data

We first compare Nelson-Aalen estimates (black lines) with the cumulative hazards obtained from a Cox model with sex as only covariate (**red lines**)



Commands for the plot on the previous slide

We start by making Nelson-Aalen plots for females and males

```
surv.s=survfit(Surv(lifetime,status==1) ~ sex, data=melanoma, stype=2)
plot(surv.s, fun="cumhaz", lty=1:2, xlim=c(0,10), ylim=c(0,0.80), main="Gender",
      xlab="Years since operation", ylab="Cumulative hazard")
legend("topleft", c("Females","Males"), lty=1:2)
```

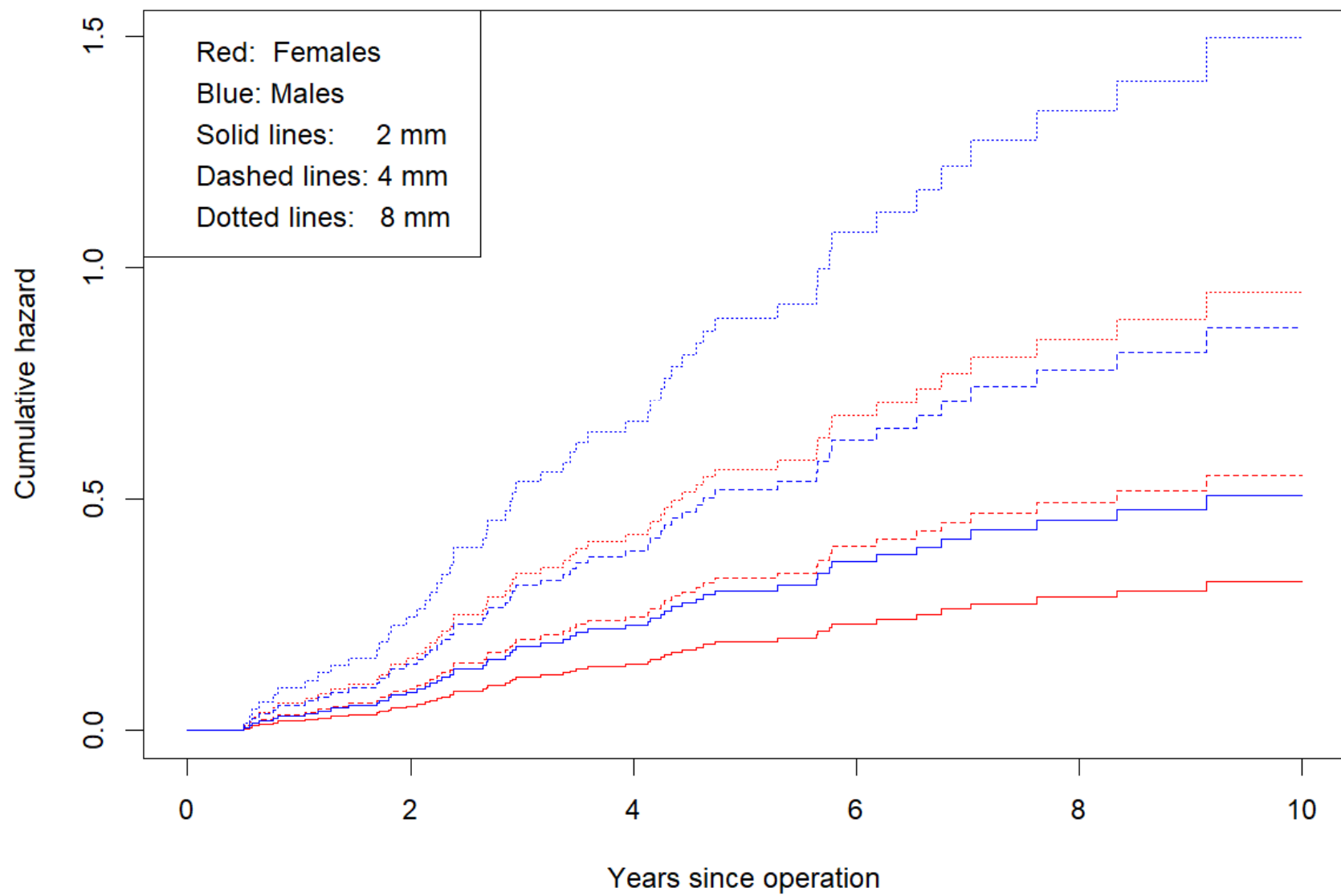
We then fit a Cox model with gender as the only covariate and plot
the model based estimates of the cumulative hazards in the same plot:

```
fit.sc=coxph(Surv(lifetime,status==1) ~ sex, data=melanoma)
surv.sc=survfit(fit.sc, newdata=data.frame(sex=c(1,2)))
lines(surv.sc, fun="cumhaz", lty=1:2, col="red")
```

Then we consider the model with sex and log-thickness

We will estimate cumulative hazards for the following combinations of the covariates

- 1) Females with thickness 2 mm
- 2) Females with thickness 4 mm
- 3) Females with thickness 8 mm
- 4) Males with thickness 2 mm
- 5) Males with thickness 4 mm
- 6) Males with thickness 8 mm



Commands for the plot on the previous slide

We consider the model with sex and log-thickness as covariates

```
fit.st=coxph(Surv(lifetime,status==1) ~sex+log2(thickn), data=melanoma)
summary(fit.st)
```

We will plot the cumulative hazards for the six covariate combinations

sex=1, thickn=2

sex=1, thickn=4

sex=1, thickn=8

sex=2, thickn=2

sex=2, thickn=4

sex=2, thickn=8

```
new.covariates=data.frame(sex=c(1,1,1,2,2,2),thickn=c(2,4,8,2,4,8))
```

```
surv.st=survfit(fit.st, newdata= new.covariates)
```

```
plot(surv.st, fun="cumhaz", xlim=c(0,10), xlab="Years since operation",
```

```
      ylab="Cumulative hazard", lty=c(1,2,3,1,2,3),
```

```
      col=c("red", "red", "red", "blue", "blue", "blue"))
```

```
legend("topleft",c("Red: Females", "Blue: Males", "Solid lines: 2 mm",
                  "Dashed lines: 4 mm", "Dotted lines: 8 mm"))
```

The survival function corresponding to an individual with a given covariate vector $\mathbf{x}_0$ is

$$S(t | \mathbf{x}_0) = \prod_{u \leq t} \{1 - dA(u | \mathbf{x}_0)\}$$

and it may be estimated by

$$\hat{S}(t | \mathbf{x}_0) = \prod_{u \leq t} \{1 - d\hat{A}(u | \mathbf{x}_0)\} = \prod_{T_j \leq t} \{1 - \Delta\hat{A}(T_j | \mathbf{x}_0)\}$$

Alternatively we may use (as is done in R):

$$\tilde{S}(t | \mathbf{x}_0) = \exp\{-\hat{A}(t | \mathbf{x}_0)\}$$

For practical purposes there is little difference between the two estimators

To plot estimated survival functions in R you just skip the "cumhaz" option in the plot command

Aalen's additive regression model

Assume that we have a sample of n individuals, and let $N_i(t)$ count the **observed** occurrences of an event of interest for individual i as a function of (study) time t

The intensity process of N_i may be given as

$$\lambda_i(t) = \underbrace{Y_i(t)}_{\text{at risk indicator}} \cdot \underbrace{\alpha(t | \mathbf{x}_i)}_{\text{hazard rate (intensity)}}$$

So far we have considered Cox's regression model

$$\alpha(t | \mathbf{x}_i) = \alpha_0(t) \exp(\boldsymbol{\beta}^T \mathbf{x}_i)$$

where the covariates have a multiplicative effect on the hazard [via the relative risk $\exp(\boldsymbol{\beta}^T \mathbf{x}_i)$]

We will now consider the non-parametric **additive** regression model (or **excess risk** regression model) due to Aalen, where the hazard rate for individual i takes the form

$$\alpha(t|\mathbf{x}_i) = \underbrace{\beta_0(t)}_{\text{baseline hazard}} + \cdots + \underbrace{\beta_p(t)x_{ip}}_{\text{excess risk at time } t \text{ per unit's increase of } x_{ip}}$$

Note that the $\beta_q(t)$'s are **regression functions**

The additive regression model is a flexible nonparametric model that allows the effect of covariates to change over time, but the model does not constrain the hazard to be non-negative

It is difficult to estimate $\beta_q(t)$ non-parametrically, so we focus on the cumulative regression functions $B_q(t) = \int_0^t \beta_q(s)ds$

At each time t we have the decomposition

$$\underbrace{dN_i(t)}_{\text{observation}} = \underbrace{\lambda_i(t)dt}_{\text{signal}} + \underbrace{dM_i(t)}_{\text{noise}}$$

Thus we at each t obtain a linear model (conditional on the past):

$$\underbrace{dN_i(t)}_{\text{observation}} = Y_i(t)dB_0(t) + \sum_{j=1}^p \underbrace{Y_i(t)x_{ij}}_{\text{covariates}} \underbrace{dB_j(t)}_{\text{parameters}} + \underbrace{dM_i(t)}_{\text{noise}}$$

We will write the model using vector and matrix notation

To this end we introduce the vectors

$$\mathbf{N}(t) = (N_1(t), N_2(t), \dots, N_n(t))^T$$

$$\mathbf{M}(t) = (M_1(t), M_2(t), \dots, M_n(t))^T$$

$$\mathbf{B}(t) = (B_0(t), B_1(t), \dots, B_p(t))^T$$

and the $n \times (p+1)$ "design matrix"

$$\mathbf{X}(t) = \begin{pmatrix} Y_1(t) & Y_1(t)x_{11} & \cdots & Y_1(t)x_{1p} \\ Y_2(t) & Y_2(t)x_{21} & \cdots & Y_2(t)x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ Y_n(t) & Y_n(t)x_{n1} & \cdots & Y_n(t)x_{np} \end{pmatrix}$$

The model on may then be written on matrix form as

$$d\mathbf{N}(t) = \mathbf{X}(t)d\mathbf{B}(t) + d\mathbf{M}(t)$$

For each time t , this is a linear regression model on matrix form (conditional on the past)

Provided $\mathbf{X}(t)$ has full rank, ordinary least squares gives

$$d\hat{\mathbf{B}}(t) = (\mathbf{X}(t)^T \mathbf{X}(t))^{-1} \mathbf{X}(t)^T d\mathbf{N}(t)$$

If we introduce the indicator

$$J(t) = I\{ \mathbf{X}(t) \text{ has full rank} \}$$

and the least squares generalized inverse

$$\mathbf{X}^-(t) = (\mathbf{X}(t)^T \mathbf{X}(t))^{-1} \mathbf{X}(t)^T$$

we obtain the estimator

$$\hat{\mathbf{B}}(t) = \int_0^t J(u) \mathbf{X}^-(u) d\mathbf{N}(u) = \sum_{T_j \leq t} J(T_j) \mathbf{X}^-(T_j) \Delta \mathbf{N}(T_j)$$

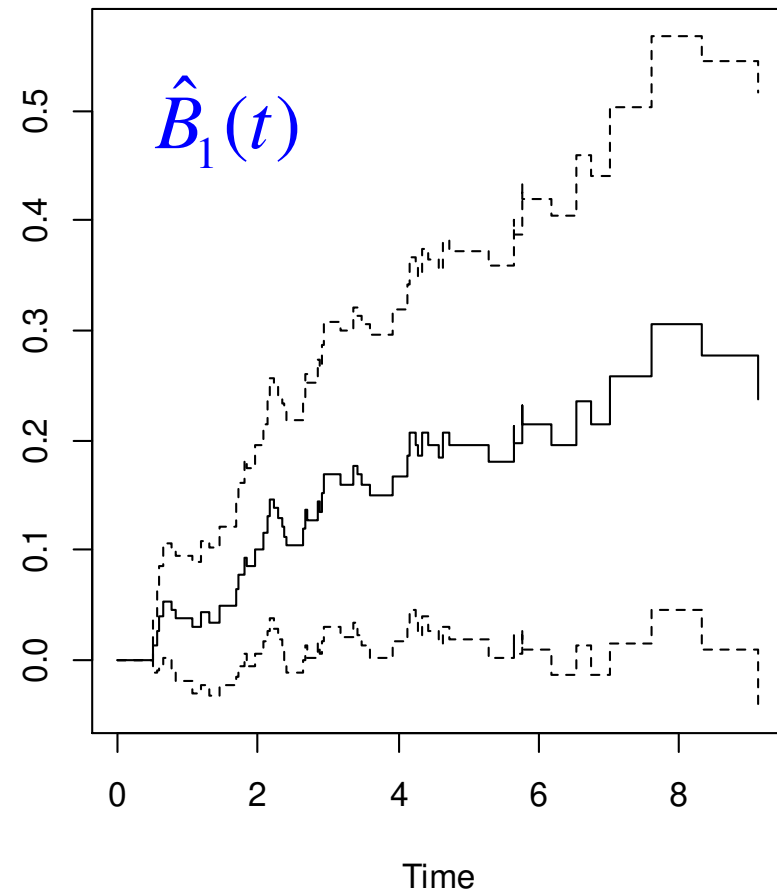
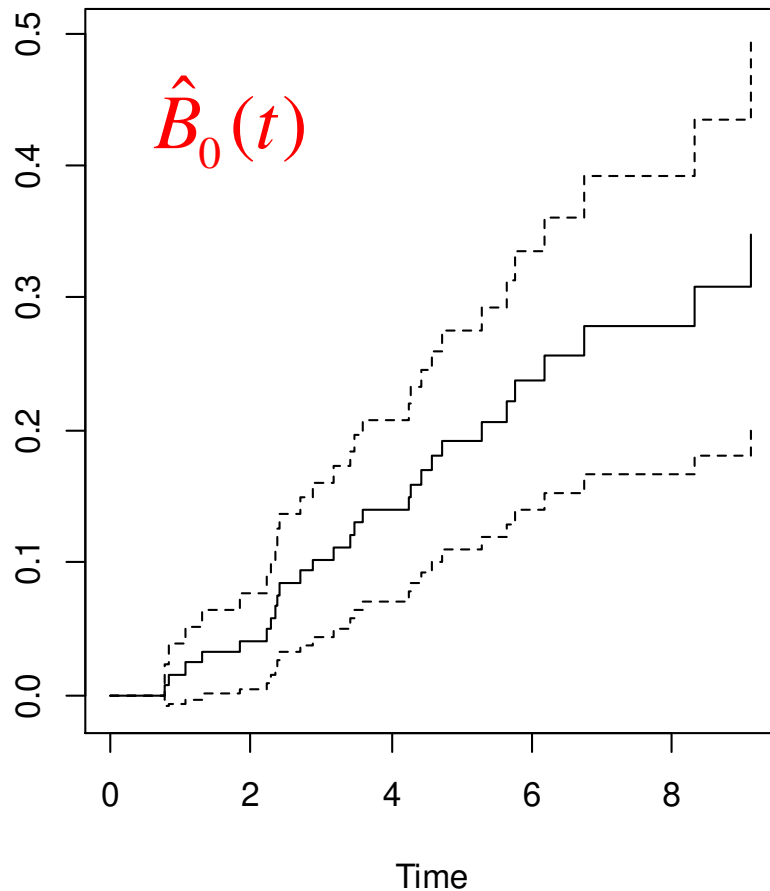
where $T_1 < T_2 < \dots$ are the event times

The estimator $\hat{\mathbf{B}}(t)$ for the additive regression model can be seen as a “multivariate Nelson-Aalen estimator”

Its statistical properties may be derived in a similar manner as for the Nelson-Aalen estimator using results on multivariate counting processes and multivariate martingales

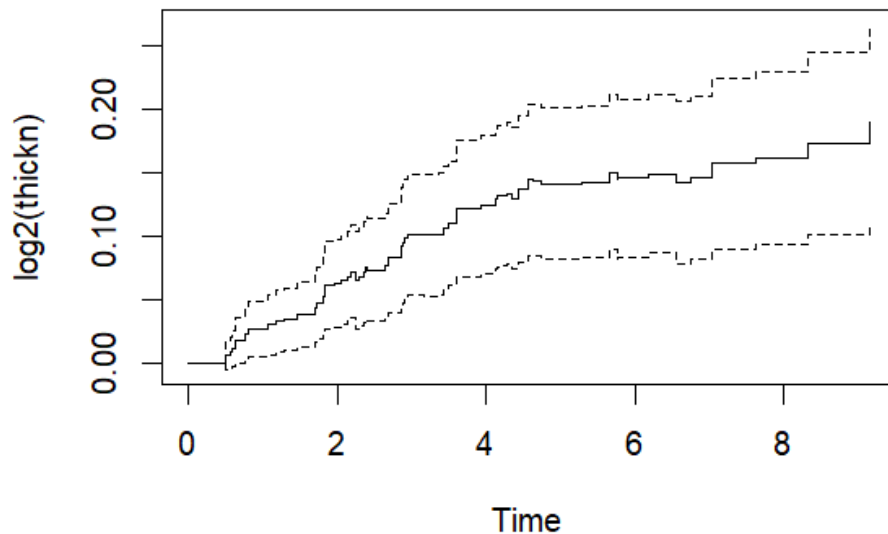
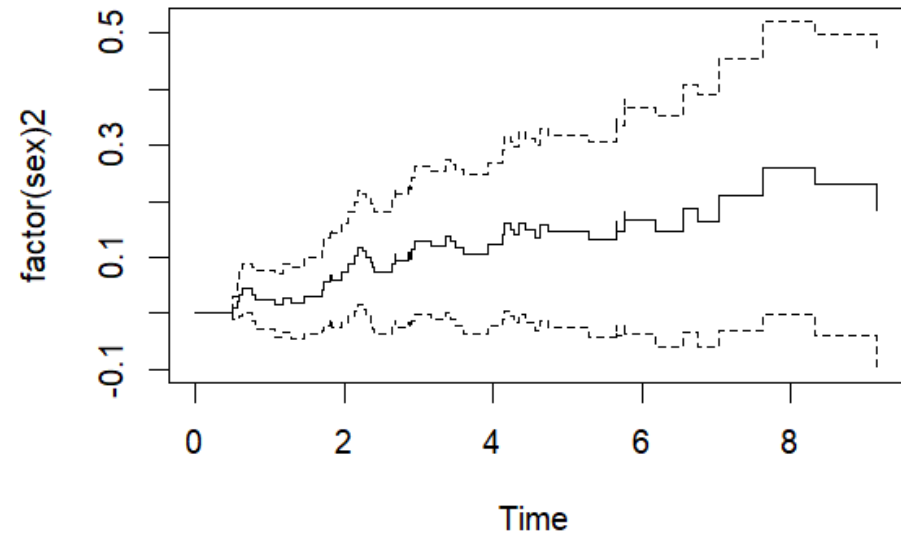
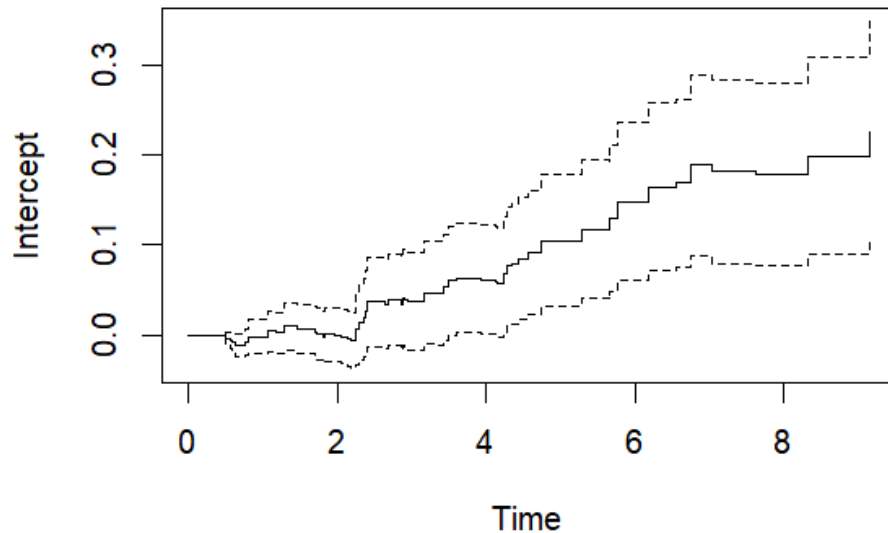
Illustration: melanoma data

Fit model with sex (as factor) as only covariate:



The **cumulative baseline** equals the Nelson-Aalen estimate for females and the **cumulative effect of sex** equals the difference between the Nelson-Aalen estimates for males and females

Fit model with sex (as factor) and log-thickness as covariates



The **cumulative baseline** corresponds to the cumulative hazard for a female with tumor thickness 1 mm

Commands for the plot on the previous slides

Read data and load survival library

```
mdata="https://www.mn.uio.no/math/english/people/aca/borgan/iwsm2026/melanoma.txt"
melanoma=read.table(mdata,header=T)
library(survival)
```

We first consider the model with sex as the only covariate:

```
fit.s=aareg(Surv(lifetime,status==1)~factor(sex), data=melanoma)
par(mfrow=c(1,2))
plot(fit.s)
```

We define sex as a factor in order to get a meaningful estimate of
the cumulative baseline hazard

Then we consider the model with sex and log-thickness:

```
fit.st=aareg(Surv(lifetime,status==1)~factor(sex)+log2(thickn), data=melanoma)
par(mfrow=c(2,2))
plot(fit.st)
```