STK4080 SURVIVAL AND EVENT HISTORY ANALYSIS

Slides 11: Regression modeling

Bo Lindqvist
Department of Mathematical Sciences
Norwegian University of Science and Technology
Trondheim

https://www.ntnu.no/ansatte/bo.lindqvist
bo.lindqvist@ntnu.no

University of Oslo, Autumn 2019

Regression models

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

$$dN_i(t) = \lambda_i(t)dt + dM_i(t)$$

We will consider regression models where the intensity process $\lambda_i(t)$ for individual i depends on a vector of (possibly) time-dependent covariates

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

The intensity for individual i may then be given as

$$\lambda_i(t) = Y_i(t)\alpha(t|\mathbf{x}_i)$$

The new issue is hence that the hazard α depends on the values of the covariates.

Regression models

A regression model specifies how the hazard rate $\alpha(t|\mathbf{x}_i)$ depends on the covariates.

We will consider two types of regression models:

- Relative risk regression models (section 4.1)
- Additive regression models (section 4.2)

A note on covariates

Throughout we will assume that the covariate processes

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

are *predictable*

This implies that:

- **fixed** covariates should be measured in advance (i.e. at time zero) and remain fixed throughout the study
- the values at time t of **time-dependent** covariates should be known "just before" time t

Covariates should not depend on information from the future!

More on covariates

It is useful to distinguish between external (or exogenous) and internal (or endogenous) covariates

Examples of external covariates are:

Fixed covariates

Defined time-dependent covariates: the covariate path is given at the outset of the study (e.g. a person's age at study time t)

Ancillary time-dependent covariates: the path of a stochastic process that is not influenced by the event being studied (e.g. observed level of air pollution)

Time-dependent covariates that are not external, are called *internal*

One example of an internal covariate is a biomarker measured for the individuals during follow-up

Interpretation of regression analyses with internal time-dependent covariates is not at all straightforward!

Relative risk regression models

Assume that the hazard rate for individual i takes the form

$$\alpha(t|\mathbf{x}_i) = \alpha_0(t) r(\boldsymbol{\beta}, \mathbf{x}_i(t))$$

We assume $r(\boldsymbol{\beta}, 0) = 1$, so the **baseline hazard** $\alpha_0(t)$ is the hazard for an individual with all covariates equal to zero.

$r(\boldsymbol{\beta}, \mathbf{x}_i(t))$ is called **the relative risk function**.

We make no assumptions of the form of the baseline hazard $\alpha_0(t)$.

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*

Cox' regression model

The common choice of relative risk function is

$$r(\boldsymbol{\beta}, \mathbf{x}_i(t)) = \exp\left(\boldsymbol{\beta}^T \mathbf{x}_i(t)\right) = \exp\left(\beta_1 x_{i1}(t) + \cdots + \beta_p x_{ip}(t)\right)$$

which gives Cox' regression model.

Consider two individuals, indexed 1 and 2, and assume that all components of $\mathbf{x}_1(t)$ and $\mathbf{x}_2(t)$ are equal, except the jth component, where $x_{2j}(t) = x_{1j}(t) + 1$.

Then:

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

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 the relative risk regression models (due to the nonparametric baseline).

Instead we have to use a partial likelihood, which we will now derive.

The intensity process of $N_i(t)$ is given as

$$\lambda_i(t) = Y_i(t)\alpha(t|\mathbf{x}_i) = Y_i(t)\alpha_0(t)r(\beta, \mathbf{x}_i(t))$$

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

$$P(dN_\bullet(t) = 1|\mathcal{F}_{t-}) = \lambda_\bullet(t) = \sum_{i=1}^n \lambda_i(t) = \sum_{i=1}^n Y_i(t)\alpha_0(t)r(\beta, \mathbf{x}_i(t))$$

Cox' partial likelihood

We consider the conditional probability of observing an event for individual i at time t, given the past and given that an event is observed at time t :

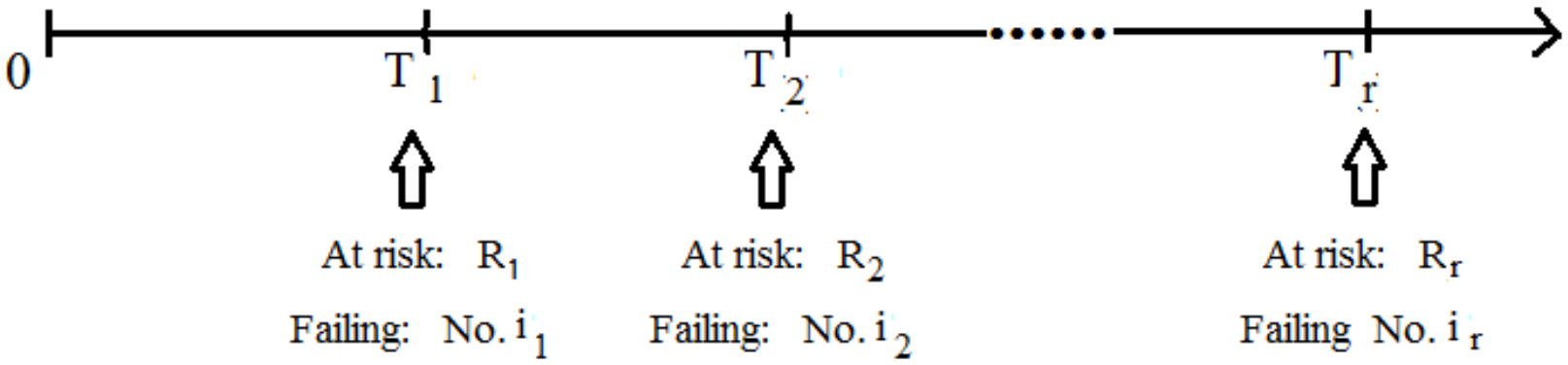
$$\begin{aligned}\pi(i|t) &= P(dN_i(t) = 1 | dN_\bullet(t) = 1, \mathcal{F}_{t-}) \\ &= \frac{P(dN_i(t) = 1|\mathcal{F}_{t-})}{P(dN_\bullet(t) = 1|\mathcal{F}_{t-})} = \frac{\lambda_i(t)}{\lambda_\bullet(t)} = \frac{Y_i(t) r(\boldsymbol{\beta}, \mathbf{x}_i(t))}{\sum_{\ell=1}^n Y_\ell(t) r(\boldsymbol{\beta},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

We obtain the **partial likelihood** for $\boldsymbol{\beta}$ by multiplying together the conditional probabilites $\pi(i|t)$ over all observed event times T_j:

$$\begin{aligned}L(\boldsymbol{\beta}) &= \prod_j \pi(i_j|T_j) = \prod_j \frac{Y_{i_j}(T_j) r(\boldsymbol{\beta}, \mathbf{x}_{i_j}(T_j))}{\sum_{\ell=1}^n Y_\ell(T_j) r(\boldsymbol{\beta}, \mathbf{x}_\ell(T_j))} \\ &= \prod_j \frac{r(\boldsymbol{\beta}, \mathbf{x}_{i_j}(T_j))}{\sum_{\ell \in \mathcal{R}_j} r(\boldsymbol{\beta}, \mathbf{x}_\ell(T_j))}\end{aligned}$$

Here i_j is the index of the individual who experiences the event at T_j, while $\mathcal{R}_j = \{\ell \mid Y_\ell(T_j) = 1\}$ is the *risk set* at T_j.

Cox' partial likelihood for β



Cox noted that since the baseline hazard $\alpha_0(t)$ is completely unknown, the times between events are not relevant for estimation of β.

Cox' *partial likelihood* is essentially the likelihood of the observed failing individuals $i_1, i_2, \ldots$:

$$L(\beta) = \text{“}P(I_1 = i_1, I_2 = i_2, \cdots, I_r = i_r)''$$

where I_j is the index of the individual that fails at time T_j.

A simple example

Model: $\alpha(t|x) = \alpha_0(t)e^{\beta x}$.

Thus we have a single fixed covariate, x, while $r(\beta, x) = e^{\beta x}$

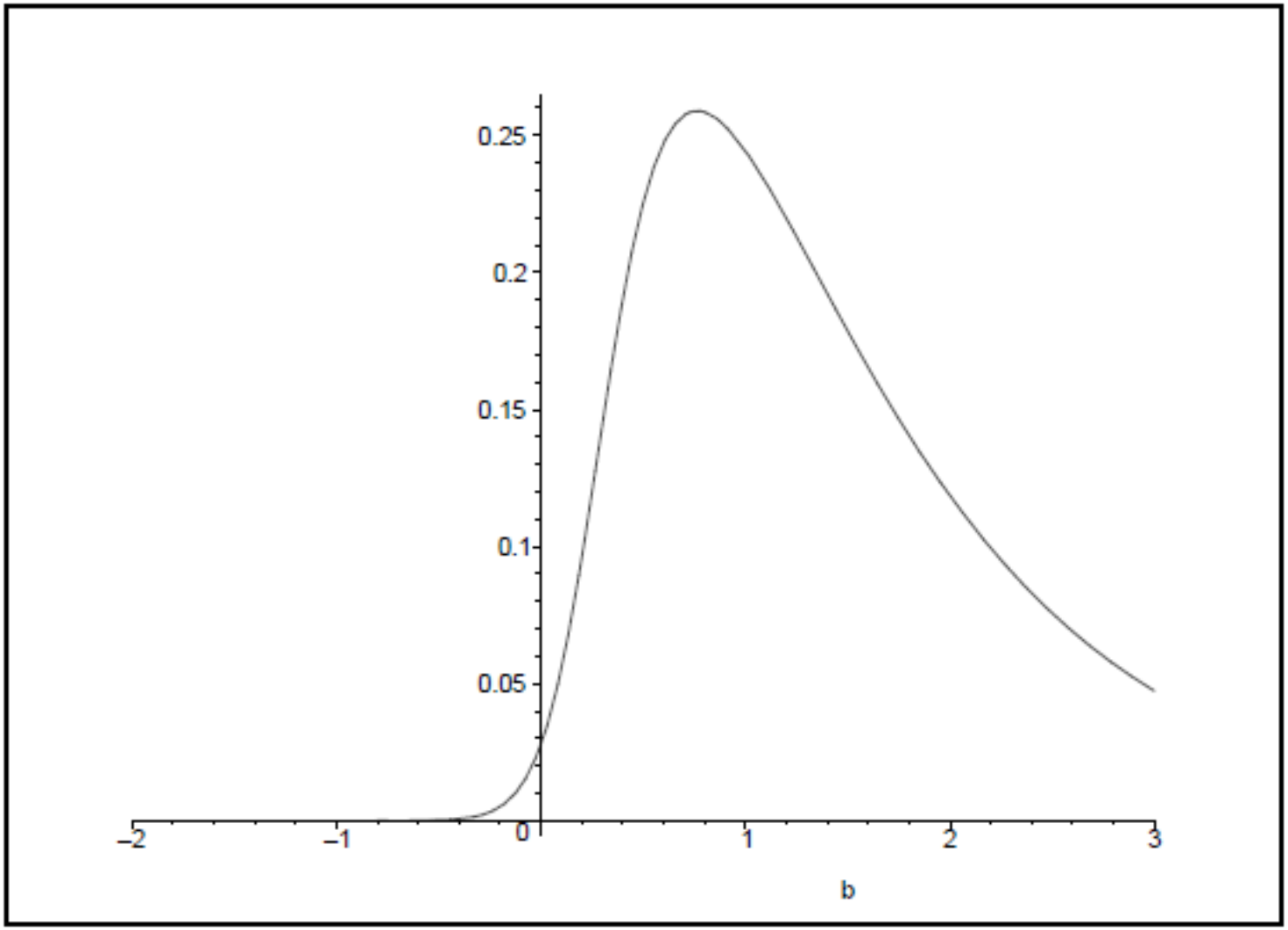
Data:

i	$\tilde{T}_i$	x_i	D_i
1	5	12	0
2	10	10	1
3	40	3	0
4	80	5	0
5	120	3	1
6	400	4	1
7	600	1	0

j	T_j	$\mathcal{R}_j$	i_j
1	10	$\{2, 3, 4, 5, 6, 7\}$	2
2	120	$\{5, 6, 7\}$	5
3	400	$\{6, 7\}$	6

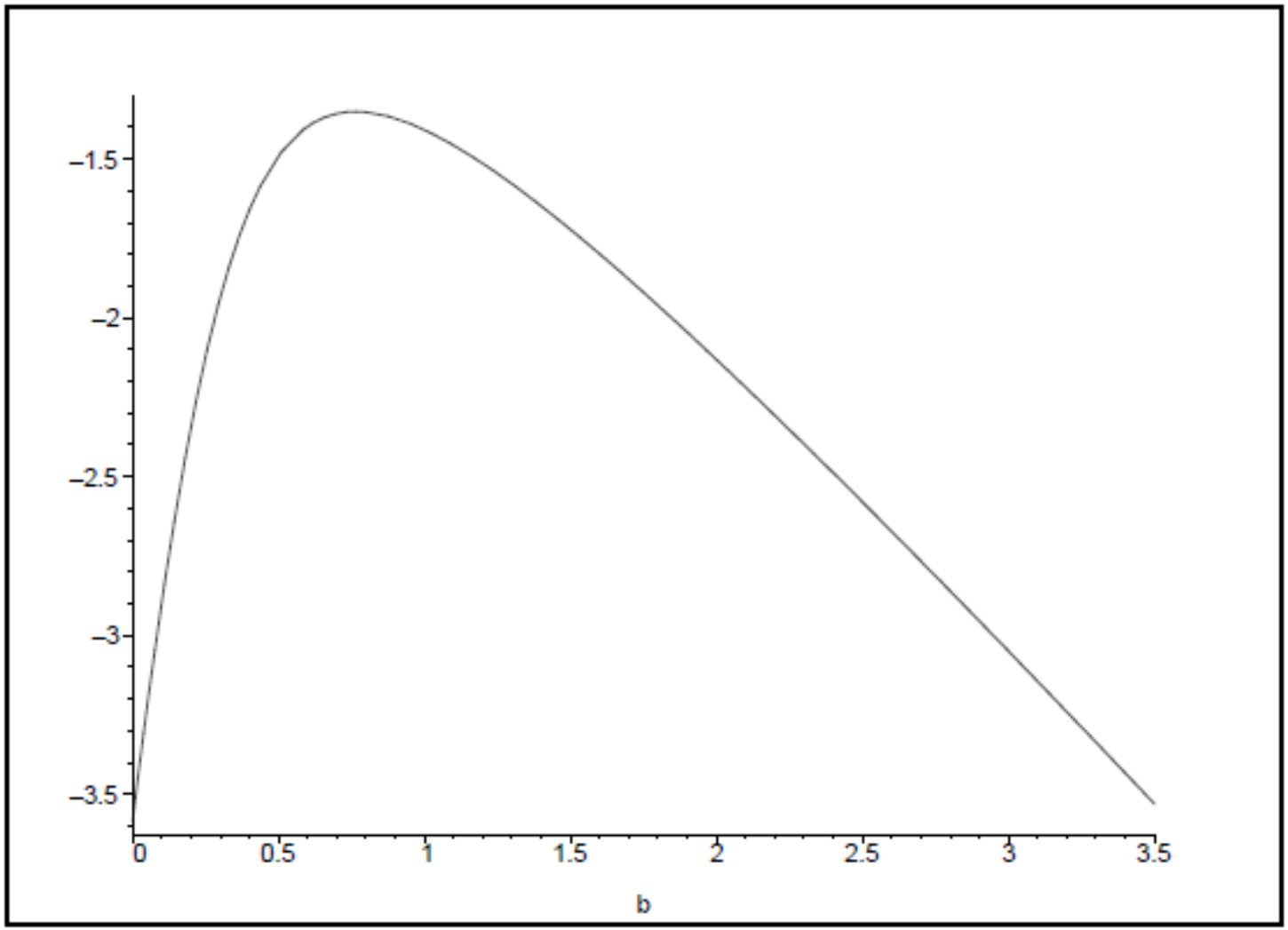
$$L(\beta) = \frac{e^{10\beta}}{e^{10\beta} + e^{3\beta} + e^{5\beta} + e^{3\beta} + e^{4\beta} + e^{\beta}} \cdot \frac{e^{3\beta}}{e^{3\beta} + e^{4\beta} + e^{\beta}} \cdot \frac{e^{4\beta}}{e^{4\beta} + e^{\beta}}$$

Simple example: Cox' partial likelihood $L(\beta)$



Maximum partial likelihood estimate: $\hat{\beta} = 0.765$.

Simple example: Cox' log partial likelihood



Maximum partial likelihood estimate: $\hat{\beta} = 0.765$.

Statistical inference in relative risk regression

It can be shown that the maximum partial likelihood estimator 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

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

is the observed information matrix.

Standard test and confidence interval for β_j

To test the null hypothesis $H_0 : \beta_j = 0$ it is common to use the Wald test statistic

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

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

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

Tests for the vector β

To test the simple null hypothesis $H_0 : \beta = \beta_0$ for a specified value of β_0 (typically 0) we may apply the usual likelihood based tests statistics:

- The likelihood ratio test statistic:

$$\chi^2_{LR} = 2\{\log L(\hat{\beta}) - \log L(\beta_0)\}$$

- The score test statistic:

$$\chi^2_{SC} = \mathbf{U}(\beta_0)^T \mathbf{I}(\beta_0)^{-1} \mathbf{U}(\beta_0)$$

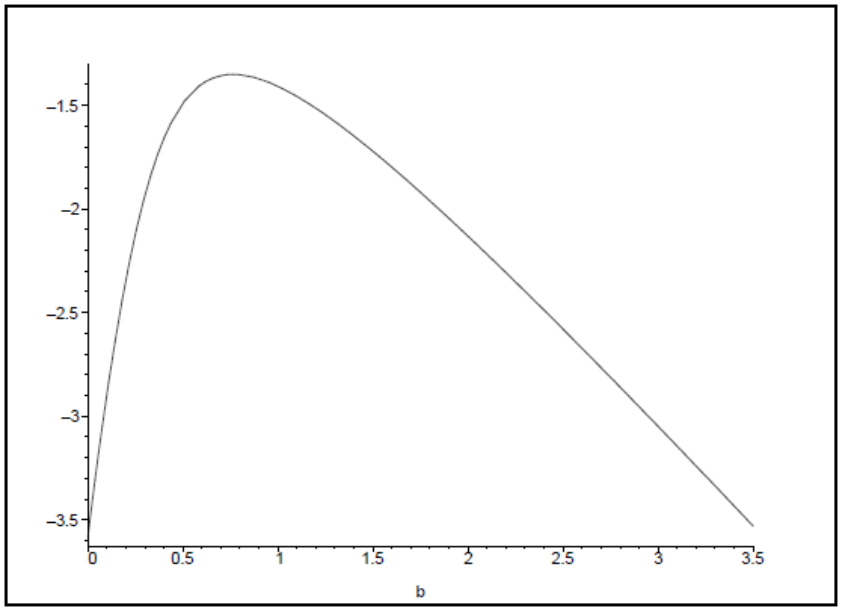
  where $\mathbf{U}(\beta) = \frac{\partial}{\partial \beta} \log L(\beta)$ is the vector of score functions

- The Wald test statistic:

$$\chi^2_W = (\hat{\beta} - \beta_0)^T \mathbf{I}(\hat{\beta})(\hat{\beta} - \beta_0)$$

All the test statistics are approximately χ^2-distributed with $df = p$ under the null hypothesis.

Simple example: Testing $H_0 : \beta = 0$



Using the data from the simple example we will test $H_0 : \beta = 0$ versus $H_1 : \beta \neq 0$ by using the likelihood ratio test:

$$\chi^2_{LR} = 2(\log L(\hat{\beta}) - \log L(0)) \sim \chi^2_1$$

under the null hypothesis.

From the figure: $\chi^2_{LR} = 2(-1.35 - (-3.45)) = 2 \cdot 2.10 = 4.2$, so we reject H_0 at 5% level (critical value 3.84).

Tests for composite hypotheses

All the tests may be generalized to a *composite null hypothesis*, where on want to test the hypothesis that r of the regression coefficients are zero (or equivalently, after a reparameterization, that there are r linear restrictions among the regression coefficients).

In particular if $\boldsymbol{\beta}^*$ is the maximum partial likelihood estimator under the null hypothesis, the likelihood ratio test statistic takes the form

$$\chi^2_{LR} = 2(\log L(\hat{\boldsymbol{\beta}}) - \log L(\boldsymbol{\beta}^*))$$

which is approximately χ^2-distributed with $df = r$ under the null hypothesis.

Using R

For illustration we use the melanoma data (cf practical exercises 1 and 2)

```
# Read data:
path="http://www.uio.no/studier/emner/matnat/math/STK4080/h14/melanoma.txt"
melanoma=read.table(path,header=T)
# We first consider the model with log-thickness as the only covariate:
fit.t=coxph(Surv(lifetime,status==1) log2(thickn),data=melanoma)
summary(fit.t)
# Note that we use base 2 logarithms for ease of interpretation
# Then we consider the model with log-thickness and sex as covariates:
fit.ts=coxph(Surv(lifetime,status==1) log2(thickn)+sex,data=melanoma)
summary(fit.ts)
# Note that since sex is a binary covariate (coded 1 and 2), we get the
# same estimates if we treat sex as a numeric covariate or as a
# categorical covariate [by using factor(sex) in the coxph-command]
# The two models may be compared using the likelihood ratio test:
anova(fit.t,fit.ts,test="Chisq")
```

Simple example with R

```
library(survival)
coxdata=read.table("https://folk.ntnu.no/bo/STK4080/cox-hand.txt",header=T)
fit.c=coxph(Surv(Time,Status==1)~x, data=coxdata)
summary(fit.c)

Call:
coxph(formula = Surv(Time, Status == 1) ~ x, data = coxdata)

  n= 7, number of events= 3

    coef exp(coef) se(coef)     z Pr(>|z|)
x 0.7650    2.1491   0.6057 1.263    0.207

  exp(coef) exp(-coef) lower .95 upper .95
x     2.149     0.4653    0.6557     7.044

Concordance= 0.875  (se = 0.242 )
Rsquare= 0.472   (max possible= 0.641 )
Likelihood ratio test= 4.46  on 1 df,   p=0.0346
Wald test            = 1.6  on 1 df,   p=0.2065
Score (logrank) test = 4.81  on 1 df,   p=0.0283
```

Estimation of cumulative baseline hazard

We will estimate the cumulative baseline hazard

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

We take the aggregated counting process $N_\bullet(t) = \sum_{i=1}^n N_i(t)$ as our starting point.

Its intensity process is given by

$$\lambda_\bullet(t) = \sum_{i=1}^n \lambda_i(t) = \left(\sum_{i=1}^n Y_i(t) r(\boldsymbol{\beta}, \mathbf{x}_i(t)) \right) \alpha_0(t)$$

If we knew $\boldsymbol{\beta}$, this would have been an example of the multiplicative intensity model.

Estimation of cumulative baseline hazard

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

$$\hat{A}_0(t;\boldsymbol{\beta}) = \int_0^t \frac{dN_\bullet(u)}{\sum_{\ell=1}^n Y_\ell(u) r(\boldsymbol{\beta}, \mathbf{x}_\ell(u))}$$

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

$$\begin{aligned}
\hat{A}_0(t;\boldsymbol{\beta}) &= \int_0^t \frac{dN_\bullet(u)}{\sum_{\ell=1}^n Y_\ell(u) r(\hat{\boldsymbol{\beta}}, \mathbf{x}_\ell(u))} \\
&= \sum_{T_j \le t} \frac{1}{\sum_{\ell \in \mathcal{R}_j} r(\hat{\boldsymbol{\beta}}, \mathbf{x}_\ell(T_j))}
\end{aligned}$$

Estimation of individual cumulative hazards

If all covariates are fixed, the cumulative hazard corresponding to an individual with a given covariate vector $\mathbf{x}_0$ is

$$A(t|\mathbf{x}_0) = \int_0^t \alpha(u|\mathbf{x}_0)du = \int_0^t r(\boldsymbol{\beta}, \mathbf{x}_0(u))\alpha_0(u)du = r(\boldsymbol{\beta}, \mathbf{x}_0)A_0(u)$$

and it may be estimated by

$$\hat{A}(t|\mathbf{x}_0) = r(\hat{\boldsymbol{\beta}}, \mathbf{x}_0)\hat{A}_0(u)$$

For a given path $\mathbf{x}_0(s) : 0 < s \leq t$ of an external time-dependent covariate, the cumulative hazard

$$A(t|\mathbf{x}_0) = \int_0^t r(\boldsymbol{\beta}, \mathbf{x}_0(u))\alpha_0(u)du$$

may be estimated by

$$\hat{A}(t|\mathbf{x}_0) = \int_0^t r(\hat{\boldsymbol{\beta}}, \mathbf{x}_0(u))d\hat{A}_0(u) = \sum_{T_j \leq t} \frac{r(\hat{\boldsymbol{\beta}}, x_0(T_j))}{\sum_{\ell \in \mathcal{R}_j} r(\hat{\boldsymbol{\beta}}, \mathbf{x}_\ell(T_j))}$$

Estimation of individual survival functions

The corresponding survival function is given by the product integral

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

and may be estimated by

$$\hat{S}(t|\mathbf{x}_0) = \prod_{u \leq t} \{1 - d\hat{A}(u|\mathbf{x}_0)\} = \prod_{T_j \leq t} \left\{ 1 - \frac{r(\hat{\boldsymbol{\beta}}, \mathbf{x}_0(T_j))}{\sum_{\ell \in \mathcal{R}_j} r(\hat{\boldsymbol{\beta}}, \mathbf{x}_\ell(T_j))} \right\}$$

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

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

The estimators of the cumulative hazards and survival functions are *approximately normal* and their variances may be estimated as described in section 4.1.6 (which is not part of the curriculum)

For illustration we continue to use the melanoma data

```
# We first consider ulceration as the only covariate and start by
# making Nelson-Aalen plots for patients with and without ulceration:
fit.su=coxph(Surv(lifetime,status==1) strata(ulcer),data=melanoma)
surv.su=survfit(fit.su)
plot(surv.su,fun="cumhaz", mark.time=F,xlim=c(0,10),ylim=c(0,0.70),
xlab="Years since operation",ylab="Cumulative hazard",lty=1:2)
legend("topleft",c("Ulceration","No ulceration"),lty=1:2)
# We then fit a Cox model with ulceration as the only covariate and plot
# the model based estmates of the cumulative hazards in the same plot:
fit.u=coxph(Surv(lifetime,status==1) ulcer,data=melanoma)
surv.u=survfit(fit.u,newdata=data.frame(ulcer=c(1,2)))
lines(surv.u,fun="cumhaz", mark.time=F,conf.int=F, lty=1:2,col="red")
```

Using R

```
# We then consider the model with ulceration and log-thickness
fit.ut=coxph(Surv(lifetime,status==1) ulcer+log2(thickn),data=melanoma)
summary(fit.ut)
# We will plot the cumulative hazards for the four covariate
combinations
# 1) ulcer=2, thickn=1
# 2) ulcer=2, thickn=4
# 3) ulcer=1, thickn=4
# 3) ulcer=1, thickn=8
new.covariates=data.frame(ulcer=c(2,2,1,1),thickn=c(1,4,4,8))
surv.ut=survfit(fit.ut,newdata= new.covariates)
plot(surv.ut,fun="cumhaz", mark.time=F, xlim=c(0,10), xlab="Years since
operation",ylab="Cumulative hazard",lty=1:4)
legend("topleft",c("1","2","3","4"), lty=1:4)
# To plot the survival functions for the same combinations of the
# covariates we just omit the "cumhaz" option:
plot(surv.ut,mark.time=F, xlim=c(0,10), xlab="Years since
peration",lty=1:4)
legend("bottomleft",c("1","2","3","4"), lty=1:4)
```

Case-study: PBC-data from Mayo Clinic

424 patients with PBC (primary biliary cirrhosis (rare disease))

A randomized clinical trial with drug DPCA versus Placebo: 312 patients chosen

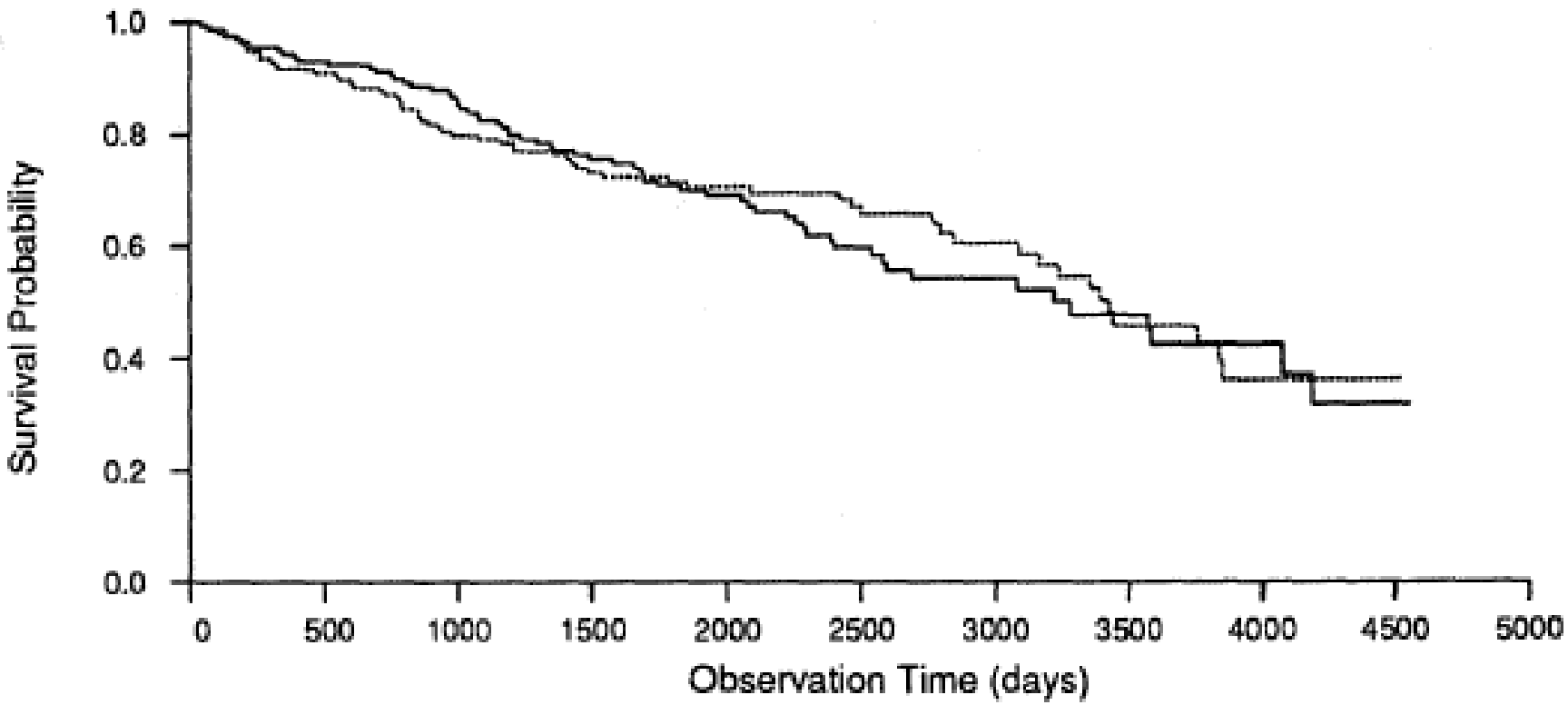
Patients included in trial: January 1974 - May 1984

Follow-up until July 1986

First: Compared DPCA group and Placebo group by Kaplan Meier.

(Source: Fleming, Thomas R., and David P. Harrington. Counting processes and survival analysis. Vol. 169. John Wiley & Sons, 2011.)

KM-plot for DPCS vs. placebo



	Time Interval				
Group	0-1000	1000-2000	2000-3000	3000-4000	4000-5000
DPCA	23/158	22/128	13/74	5/31	2/10
Placebo	31/154	12/120	7/70	10/32	0/11

(# events/# at risk)

Figure 4.4.1 Estimated survival curves in DPCA and placebo groups, PBC data.

Cox regression model for DPCS vs. placebo

Model: $\alpha(t|x) = \alpha_0(t)e^{\beta x}$

x=0 for DCPA $\alpha_0(t)$
x=1 for Placebo $\alpha_0(t)e^{\beta}$
$\hat{\beta} = -0.0571$, $\chi^2_{LR} = 2(\log L(\hat{\beta}) - \log L(0)) = 0.102$ (not significant)
$\widehat{SE(\hat{\beta})} = 0.1792$
95% confidence interval for β : $\hat{\beta} \pm 1.96 \cdot 0.1792$
(-0.408, 0.294)

so CI for relative risk e^{β}: (0.66, 1.34)

Conclusion: In the best case the new drug leads to 1.34 relative risk for not using it (would need at least 1.50 to do further investigations).

Natural history model for PBC

The data on the 312 PBC randomized patients can be used to build a statistical model for the influence of covariates on disease outcome.

The data contains 14 clinical, biochemical and histological variables.

Their model is (now $\alpha(\cdot)$ is used instead of $z(\cdot)$ for hazard rate):

$$\alpha(t|\mathbf{x}) = \alpha_0(t)e^{\beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k}$$

In the beginning k=14

Covariates

Table 4.4.1 Prognostic Factors: Summary of Univariate Statistics (312 Patients in the PBC Clinical Trial of DPCA)

Demographic	min	1st Q	med	3rd Q	max	Missing	Rao χ^2(1 d.f.)
Age (years)	26.3	42.1	49.8	56.7	78.4	0	20.86
Sex	male:	36	female:	276		0	4.27

Clinical		Absent		Present		Missing	Rao χ^2(1 d.f.)
Ascites		288		24		0	104.02
Hepatomegaly		152		160		0	40.18
Spiders		222		90		0	30.31
Edema[1]	0: 263	1/2: 29		1: 20		0	97.89

Biochemical	min	1st Q	med	3rd Q	max	Missing	Rao χ^2(1 d.f.)
Bilirubin	0.3	0.8	1.35	3.45	28.0	0	190.62
Albumin	1.96	3.31	3.55	3.80	4.64	0	70.83
Urine Copper	4	41	73	123	588	2	84.35
Pro Time	9.0	10.0	10.6	11.1	17.1	0	51.76
Platelet Count	62	200	257	323	563	4	12.15
Alkaline Phos	289	867	1259	1985	13862	0	2.58
SGOT	26	81	115	152	457	0	29.59

Histologic	1	2	3	4	Missing	Rao χ^2(1 d.f.)
Stage	16	67	120	109	0	46.49

Which covariates to keep in the model?

$\rightarrow$ Bilirubin most significant

$\rightarrow$ Take out expensive/complicated covariates:
stage, urine, copper, SGOT

Remains 11 variables; then a step-down procedure is used to eliminate one (non-significant) variable at a time, arriving at lower table on next slide.

Variable selection: Table

Table 4.4.2 Results of variable selection procedure in 312 randomized cases with PBC.

(a) First Step, log likelihood −550.603

	Coef.	Std. Err.	Z stat.
Age	2.819 e-2	9.538 e-3	2.96
Albumin	−9.713 e-1	2.681 e-1	−3.62
Alk. Phos	1.445 e-5	3.544 e-5	0.41
Ascites	2.813 e-1	3.093 e-1	0.91
Bilirubin	1.057 e-1	1.667 e-2	6.34
Edema	6.915 e-1	3.226 e-1	2.14
Hepatomegaly	4.853 e-1	2.913 e-1	2.21
Platelets	−6.063 e-4	1.025 e-3	−0.59
Prothrombin Time	2.428 e-1	8.420 e-2	2.88
Sex	−4.769 e-1	2.643 e-1	−1.80
Spiders	2.889 e-1	2.093 e-1	1.38

(b) Last Step, log likelihood −554.237

	Coef.	Std. Err.	Z stat.
Age	0.0338	0.00925	3.65
Albumin	−1.0752	0.24103	−4.46
Bilirubin	0.1070	0.01528	7.00
Edema	0.8072	0.30775	2.62
Hepatomegaly	0.5903	0.21179	2.79
Prothrombin Time	0.2603	0.07786	3.34

Variable selection: Explanation

Table 4.4.2: Cox with 11 variable.

Recall: *Z stat* means Coef/Std.Err.

Step-down procedure: From (a) to (b): 5 variables taken out;

Log-likelihood statistic:

$$2 \cdot \text{ difference in log likelihood } = 7.268$$

should be compared to χ^2_5 : $P(\chi^2_5 > 7.268) = 0.201$, so we do not reject the null hypothesis that all these 5 variables have coefficients equal to 0.

log-transformations

Then is considered log-transformations of continuous variables - four variables using logs are added to model, and this leads to increased likelihood!

Finally: Arrives at model 4.4.3(c)

Final model (c)

Table 4.4.3 Regression models with log transformations of continuous variables, 312 randomized cases with PBC.

(a) Log likelihood −538.274

	Coef.	Std. Err.	Z stat.
Age	−0.0289	0.07141	−0.41
log(age)	3.2248	3.71828	0.87
Albumin	1.0068	1.73450	0.58
log(Albumin)	−5.8629	5.42315	−1.08
Bilirubin	−0.0461	0.03547	−1.30
log(Bilirubin)	1.0774	0.21127	5.10
Edema	0.8238	0.30386	2.71
Prothrombin Time	−0.6175	1.14523	−0.54
log(Pro Time)	10.1928	13.36131	0.76
Hepatomegaly	0.1964	0.22628	0.87

(b) Log likelihood −541.064

	Coef.	Std. Err.	Z stat.
Age	0.0337	0.00864	3.89
Albumin	−0.9473	0.23713	−3.99
log(Bilirubin)	0.8845	0.09854	8.98
Edema	0.8006	0.29914	2.68
Prothrombin Time	0.2463	0.08426	2.92

(c) Log likelihood −540.412

	Coef.	Std. Err.	Z stat.
Age	0.0333	0.00866	3.84
log(Albumin)	−3.0553	0.72408	−4.22
log(Bilirubin)	0.8792	0.09873	8.90
Edema	0.7847	0.29913	2.62
log(Prothrombin Time)	3.0157	1.02380	2.95

Estimation of survival probabilities

Recall:

$$S(t|\mathbf{x}) = \exp\{-A(t|\mathbf{x})\} = \exp\{-A_0(t)e^{\boldsymbol{\beta}'\mathbf{x}}\} = \exp\{-A_0(t)e^{R}\}$$

where $R = \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k = \boldsymbol{\beta}'\mathbf{x}$ is called *Risk Score*.

The estimated survival function for a patient with estimated risk score $\hat{R}$ is hence

$$\hat{S}(t|\hat{R}) = e^{-\hat{A}_0(t)e^{\hat{R}}}$$

In the data we have the median risk score $\hat{R} = 5.24$, and for this value we get the one- and five-year survival estimates:

$$\hat{S}(1|\hat{R}) = 0.982, \quad \hat{S}(5|\hat{R}) = 0.845$$

A low-risk example: Age 52; Albumin 4.5; Bilirubin 0.5; Edema 0; Prothrombin 10.1; gives

$$\hat{R} = 0.0333 \cdot 52 - 3.0553 \cdot \ln 4.5 + 0.879 \cdot \ln 0.5 - \cdots = 3.49$$

so $\hat{S}(5|\hat{R}) = 0.97$