STK4080/9080 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.edu/employees/bo.lindqvist
bo.lindqvist@ntnu.no
boli@math.uio.no

University of Oslo, Spring 2021

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, for example **Cox models**)
- Additive regression models (section 4.2, **Aalen's additive model**))

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 the covariate is constant throuhgout the study

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**. An internal covariate is typically the output of a stochastic process generated by the individual under study and observed only as long as the subject survives and uncensored.

Special care is needed in the handling and interpretation of **internal** *time-dependent covariates. We will mainly treat the case of external covariates.*

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 maximum likelihood estimation (ML) 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 :

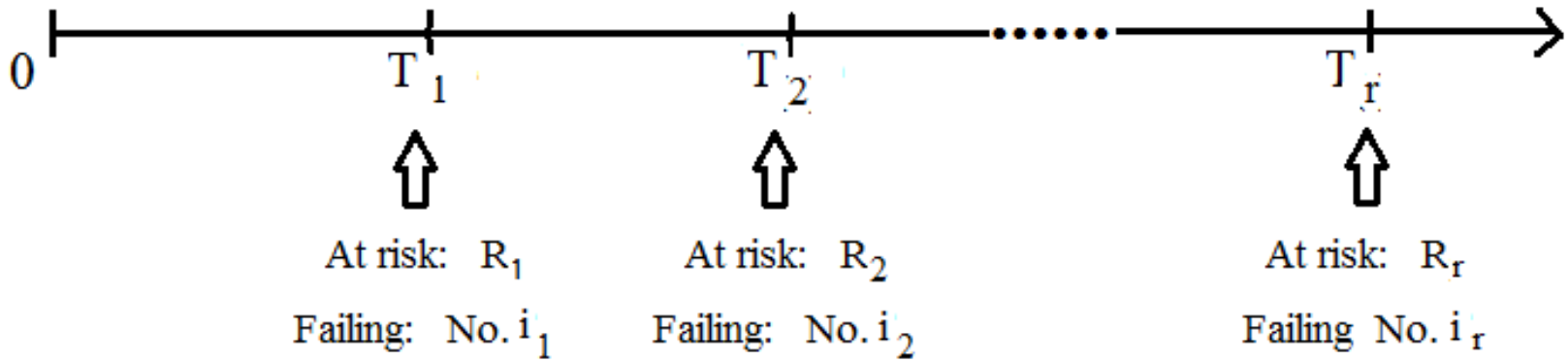
$$\begin{aligned}\pi(i|t) &= P(dN_i(t) = 1 | dN_\bullet(t) = 1, \mathcal{F}_{t-}) \\ &= \frac{P(dN_i(t) = 1|\mathcal{F}_{t-})}{P(dN_\bullet(t) = 1|\mathcal{F}_{t-})} = \frac{\lambda_i(t)}{\lambda_\bullet(t)} = \frac{Y_i(t) r(\boldsymbol{\beta}, \mathbf{x}_i(t))}{\sum_{\ell=1}^n Y_\ell(t) r(\boldsymbol{\beta},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

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 $r(\beta, x) = e^{\beta x}$

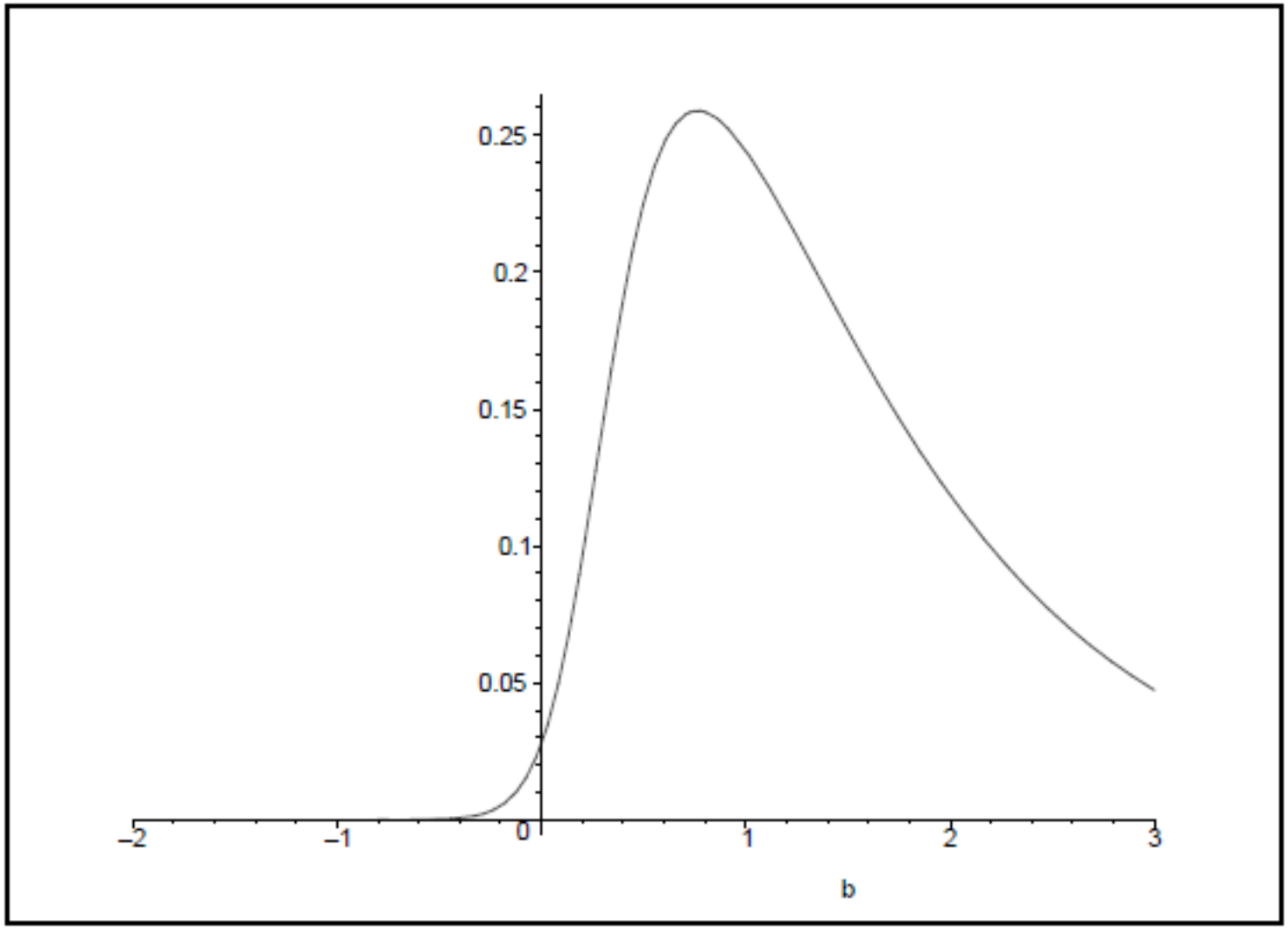
Data:

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 for $D_j = 1$	$\mathcal{R}_j$	i_j
1	10	$\{2, 3, 4, 5, 6, 7\}$	2
2	120	$\{5, 6, 7\}$	5
3	400	$\{6, 7\}$	6

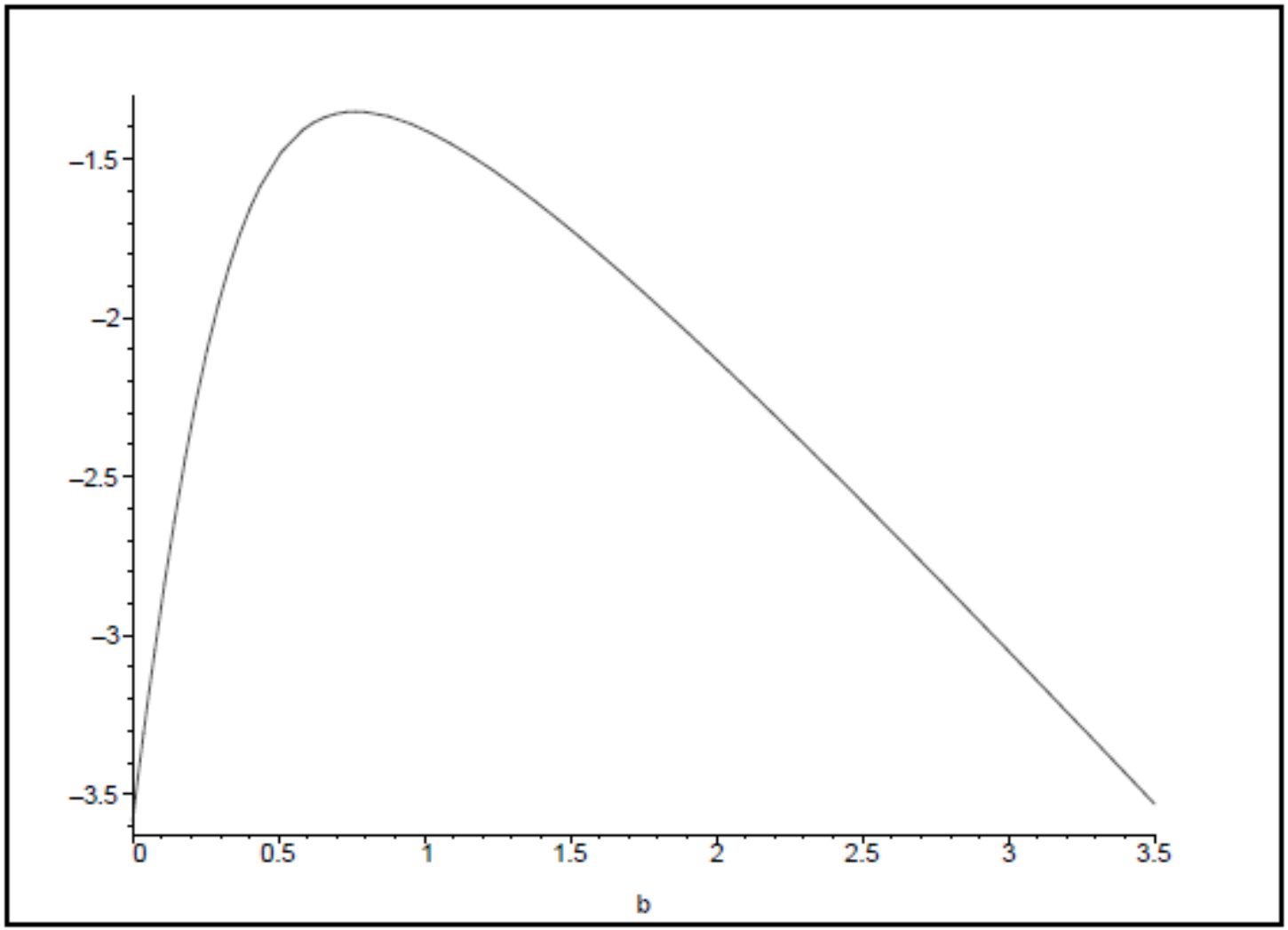
$$L(\beta) = \frac{e^{10\beta}}{e^{10\beta} + e^{3\beta} + e^{5\beta} + e^{3\beta} + e^{4\beta} + e^{\beta}} \cdot \frac{e^{3\beta}}{e^{3\beta} + e^{4\beta} + e^{\beta}} \cdot \frac{e^{4\beta}}{e^{4\beta} + e^{\beta}}$$

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.

(We will have a closer look at the asymptotic properties of $\hat{\boldsymbol{\beta}}$ in Slides 12, or see Section 4.1.5 in ABG).

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 $\boldsymbol{\beta}$

To test the simple null hypothesis $H_0 : \boldsymbol{\beta} = \boldsymbol{\beta}_0$ for a specified value of $\boldsymbol{\beta}_0$ (typically 0) we may apply the usual likelihood based tests statistics:

- The likelihood ratio test statistic:

$$\chi^2_{LR} = 2\{\log L(\hat{\boldsymbol{\beta}}) - \log L(\boldsymbol{\beta}_0)\}$$

- The score test statistic:

$$\chi^2_{SC} = \mathbf{U}(\boldsymbol{\beta}_0)^T \mathbf{I}(\boldsymbol{\beta}_0)^{-1} \mathbf{U}(\boldsymbol{\beta}_0)$$

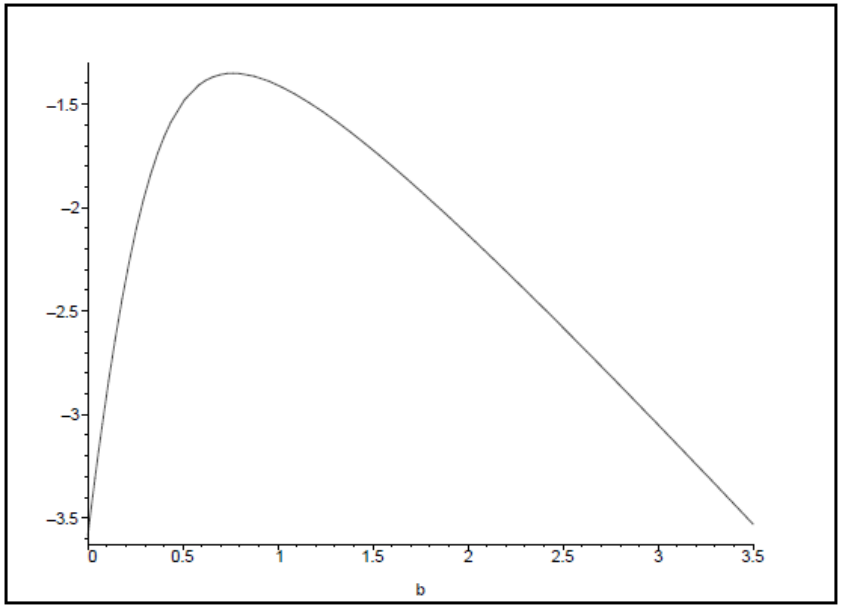
  where $\mathbf{U}(\boldsymbol{\beta}) = \frac{\partial}{\partial \boldsymbol{\beta}} \log L(\boldsymbol{\beta})$ is the vector of score functions

- The Wald test statistic:

$$\chi^2_{W} = (\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)^T \mathbf{I}(\hat{\boldsymbol{\beta}})(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)$$

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.

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 = 0.145 )
Likelihood ratio test= 4.46  on 1 df,   p=0.03
Wald test            = 1.6  on 1 df,   p=0.2
Score (logrank) test = 4.81  on 1 df,   p=0.03
```

Example: Cox-regression in melanoma-data using R

For information on these data, see *Tutorial for the Nelson-Aalen estimator*,

```
# 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)~logthick,data=melanoma)
summary(fit.t)
# Then we consider the model with log-thickness and sex as covariates:
fit.ts=coxph(Surv(lifetime,status==1)~logthick+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")
```

Example: Cox-regression in melanoma-data using R

```
> fit.ts=coxph(Surv(lifetime,status==1)~logthick+sex,data=melanoma)
> summary(fit.ts)
Call:
coxph(formula = Surv(lifetime, status == 1) ~ logthick + sex,
    data = melanoma)

  n= 205, number of events= 57

            coef exp(coef) se(coef)     z Pr(>|z|)
logthick 0.7809    2.1834   0.1573 4.963 6.94e-07 ***
sex      0.4580    1.5809   0.2687 1.705   0.0883 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

         exp(coef) exp(-coef) lower .95 upper .95
logthick     2.183     0.4580    1.6040     2.972
sex          1.581     0.6326    0.9337     2.677

Concordance= 0.749  (se = 0.033 )
Likelihood ratio test= 33.45  on 2 df,   p=5e-08
Wald test            = 31  on 2 df,   p=2e-07
Score (logrank) test = 32.52  on 2 df,   p=9e-08
```

Estimation of cumulative baseline hazard

For complete estimation of a survival regression model, we will need to estimate the baseline hazard. As for the non-regression case, we will instead estimate the cumulative 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

Assuming first that $\boldsymbol{\beta}$ is known, 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 simply replace it by $\hat{\boldsymbol{\beta}}$, to obtain

The Breslow estimator:

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

Estimation of individual cumulative hazards

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

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

which 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)

Example with melanoma-data using R

Consider now the binary covariate `ulcer` (*ulceration*) as the only covariate.

Let us first make *separate* Nelson-Aalen plots for patients with and without ulceration by using `strata` as follows:

```
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 may then fit a Cox model $\alpha_0(t)e^{\beta x}$ with $x =$`ulcer` as the only covariate, and plot the model based estimates of the cumulative hazards in the same plot as above:

```
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")
```

...melanoma-data using R

Cumulative hazards estimated by (black:) separate NA plots; (red:) estimated Cox model.

...melanoma-data using R

```
# We then consider the model with covariates ulcer and logthick
fit.ut=coxph(Surv(lifetime,status==1)~ulcer+log(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
surgery",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
surgery",ylab="Survival probability",lty=1:4)
legend("bottomleft",c("1","2","3","4"), lty=1:4)
```

...melanoma-data using R