

# Department of Natural Sciences

<http://www.dina.kvl.dk/~torbenm/DINA/survival>

## Cox-regression part II

Torben Martinussen

[torbenm@dina.kvl.dk](mailto:torbenm@dina.kvl.dk)

# Outline

- Cox-regression
- Checking Assumptions

# Cox-Regression: checking assumptions

Wish to study a treatment effect while correcting for other variables. May use Cox-model:

$$\lambda_i(t) = \lambda_0(t) \exp(\beta_1 X_{i1} + \dots + \beta_p X_{ip})$$

$\lambda_0(t)$  is the baseline hazard for a subject with covariates 0.

The Cox models ability to deal with many covariates comes from the regression structure.

Some assumptions have been made

- The effects of covariates are additive and linear on the log risk scale.
- If covariates interact with each other the regression model should include interaction terms.
- The relative risk between the hazard rate for two subjects is constant over time

$$c(\beta_1, \dots, \beta_p) = \frac{\lambda'_i(t)}{\lambda_i(t)}$$

In other words, there is no time interaction with the covariates.

# Investigating Interactions

- Interaction is an important statistical concept
- It means that the **effect (on survival) of a covariate may change according to the value of another covariate**

```
> fit1<-coxph(Surv(time,status==1)~treatF+treatI+age.gr1+age.gr2)
> fit1
Call:
coxph(formula = Surv(time, status == 1) ~ treatF + treatI + age.gr1 +
      age.gr2)
```

	coef	exp(coef)	se(coef)	z	p
treatF	0.123	1.13	0.0885	1.39	1.6e-01
treatI	0.604	1.83	0.0881	6.86	7.1e-12
age.gr1	0.829	2.29	0.0941	8.81	0.0e+00
age.gr2	0.426	1.53	0.0912	4.67	3.1e-06

```
Likelihood ratio test=126 on 4 df, p=0 n= 894
```

Let us see whether the effect of the treatments changes in the age groups:

```
> fit1<-coxph(Surv(time,status==1)~factor(age.gr)*factor(treat))
```

with result

# Interactions

```
> fit1
```

```
Call:
```

```
coxph(formula = Surv(time, status == 1) ~ factor(age.gr) * factor(treat))
```

	coef	exp(coef)	se(coef)	z	p
factor(age.gr)1	1.361	3.900	0.172	7.916	2.4e-15
factor(age.gr)2	0.728	2.071	0.171	4.261	2.0e-05
factor(treat)1	0.705	2.023	0.181	3.904	9.4e-05
factor(treat)2	0.939	2.557	0.182	5.157	2.5e-07
factor(age.gr)1:factor(treat)1	-0.819	0.441	0.229	-3.572	3.5e-04
factor(age.gr)2:factor(treat)1	-0.745	0.475	0.232	-3.204	1.4e-03
factor(age.gr)1:factor(treat)2	-0.768	0.464	0.239	-3.210	1.3e-03
factor(age.gr)2:factor(treat)2	-0.178	0.837	0.226	-0.788	4.3e-01

```
Likelihood ratio test=149 on 8 df, p=0 n= 894
```

- Indication of interaction?

# Interaction between categorical and cont. variables

Melanoma data:

Consider the Cox-model for melanoma data with explanatory variables sex and log(thickness) (lt):

$$\lambda_i(t) = \lambda_0(t) \exp(\beta_1 \cdot \text{sex}_i + \beta_2 \cdot \text{lt}_i)$$

How do we check for interaction in this situation?

- Group the continuous variable (log(thickness)) and proceed as before.
- Allow for two different  $\beta_2$ 's, one for each value of sex.

# Interaction between categorical and cont. variables

In R:

```
> lthick<-log(thickness)
> fit1<-coxph(Surv(time,status==1)~sex+lthick+sex*lthick)
> fit1
```

	coef	exp(coef)	se(coef)	z	p
sex	0.591	1.807	0.454	1.304	1.9e-01
lthick	0.834	2.302	0.214	3.895	9.8e-05
sex:lthick	-0.113	0.893	0.311	-0.363	7.2e-01

Cox-model:  $\lambda_0(t) \exp(\beta_1 \cdot \text{sex} + \beta_2 \cdot \text{lt} + \beta_3 \cdot \text{sexlt})$

with

$$\beta_1 \cdot \text{sex} + \beta_2 \cdot \text{lt} + \beta_3 \cdot \text{sexlt} = \begin{cases} \beta_1 + (\beta_2 + \beta_3) \cdot \text{lt}, & \text{sex} = 1 \\ \beta_2 \cdot \text{lt}, & \text{sex} = 0 \end{cases}$$

- The baseline  $\lambda_0(t)$  is the hazard for?
- The effect of  $\text{lt}$  is estimated to ? for male and to ? for female
- Is there an interaction between  $\text{sex}$  and  $\text{lt}$ ?

# Linearity of covariate effects

On the log-scale the intensity depends linearly of the covariates according to the Cox-model

$$\log(\lambda_i(t)) = \log(\lambda_0(t)) + \beta_1 X_{i1} + \dots + \beta_p X_{ip}.$$

Check for linearity of the effect is done just like for any other regression model.

We need to see if the model provides a reasonable description of the data:

- (i) The effect of a variable is linear  $\beta_1 X_{i1}$ .
- (ii) The effect of different covariates are additive (interactions).

When all covariates are class variables (i) is not necessary.

To see if (ii) is satisfied, one must check if there is a significant interaction between the covariates. What interactions to investigate, should be guided by subject matter and biological knowledge.

# Wrom-data

Consider the continuous variable age, which gives the age of the parent worms

To see if (i) is satisfied: **Either**: one can compare a simple linear fit of  $X_{i1}$  with

$$\log(\lambda_i(t)) = \log(\lambda_0(t)) + \beta_1 \text{age}_i + \beta_2 \text{age}_i^2$$

```
> age2<-age*age
> fit<-coxph(Surv(time,status==1)~age)
> fit2<-coxph(Surv(time,status==1)~age+age2)
> fit2
              coef exp(coef) se(coef)      z      p
age  -0.29168      0.747  0.07996 -3.65 0.00026
age2  0.00765      1.008  0.00320  2.39 0.01700
Likelihood ratio test=105  on 2 df, p=0  n= 894
> fit
              coef exp(coef) se(coef)      z p
age -0.102      0.903  0.0104 -9.84 0
Likelihood ratio test=99.5  on 1 df, p=0  n= 894
```

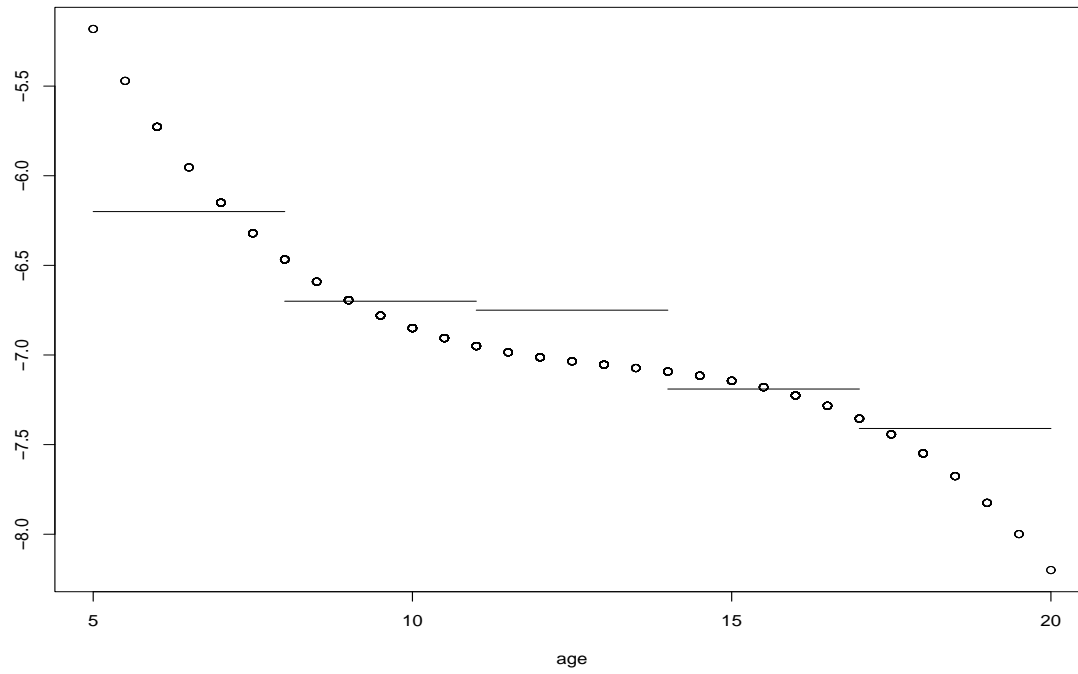
The effect of the quadratic term may be tested using Wald-test or likelihood ratio test.

# Example continued

**Or:** We start by categorizing the age into 5 groups say, and then estimate the effect for each group to get a feel for what the data really tell us.

Categorize age into groups (5-8,8-11,11-14,14-17,17-20) and then fit a model with a class variable for these groups.

```
> age.c1<-0*age+1*(age<8);
> age.c2<-0*age+1*( (age>=8)&(age<11) );
> age.c3<-0*age+1*( (age>=11)&(age<14) );
> age.c4<-0*age+1*( (age>=14)&(age<17) );
> age.c5<-0*age+1*( (age>=17) );
>
> fitgr<-coxph(Surv(time,status==1)~age.c2+
+             age.c3+age.c4+age.c5)
> fitgr
      coef exp(coef) se(coef)      z      p
age.c2 -0.503      0.604   0.118 -4.26 2.1e-05
age.c3 -0.549      0.578   0.124 -4.44 9.0e-06
age.c4 -0.988      0.372   0.128 -7.70 1.4e-14
age.c5 -1.209      0.298   0.147 -8.23 2.2e-16
> fit3<-coxph(Surv(time,status==1)~age+age2+age3)
> fit3
      coef exp(coef) se(coef)      z      p
age  -1.53484      0.215 0.399307 -3.84 0.00012
age2  0.11335      1.120 0.033629  3.37 0.00075
age3 -0.00285      0.997 0.000906 -3.14 0.00170
```



# Checking proportionality

We finally turn attention to the assumption of constant relative risk between the hazard rates for two subjects

$$c(\beta_1, \dots, \beta_p) = \frac{\lambda'_i(t)}{\lambda_i(t)}$$

We start by considering the 2-sample case. In the 2-sample case the assumption is equivalent to proportional intensities for the two groups, i.e.,

$$\exp(\beta_1)\lambda_1(t) = \lambda_2(t).$$

This implies that the cumulative intensities are proportional

$$\Lambda_2(t) = \int_0^t \lambda_2(s)ds = \exp(\beta_1)\Lambda_1(t)$$

So a plot of  $\hat{\Lambda}_2(t)$  versus  $\hat{\Lambda}_1(t)$  should give a straight line through (0,0) with slope  $\exp(\beta_1)$ . Similarly

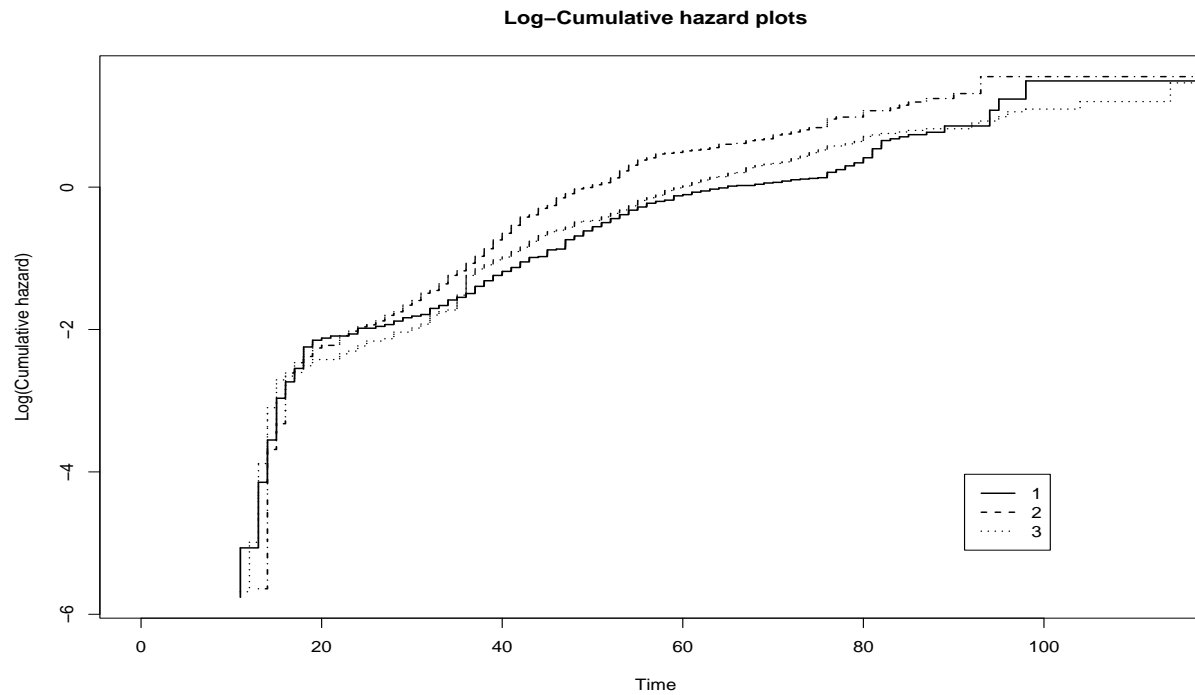
$$\log(\Lambda_2(t)) - \log(\Lambda_1(t)) = \beta_1$$

so plotting  $\log(\hat{\Lambda}_k(t))$ ,  $k = 1, 2$ , versus  $t$  should give parallel curves.

# Checking Proportionality

For the worm-data we consider if treatment effects are proportional. In R:

```
> fit<-coxph(Surv(time,status==1)~strata(treat))  
> plotkum(fit)
```

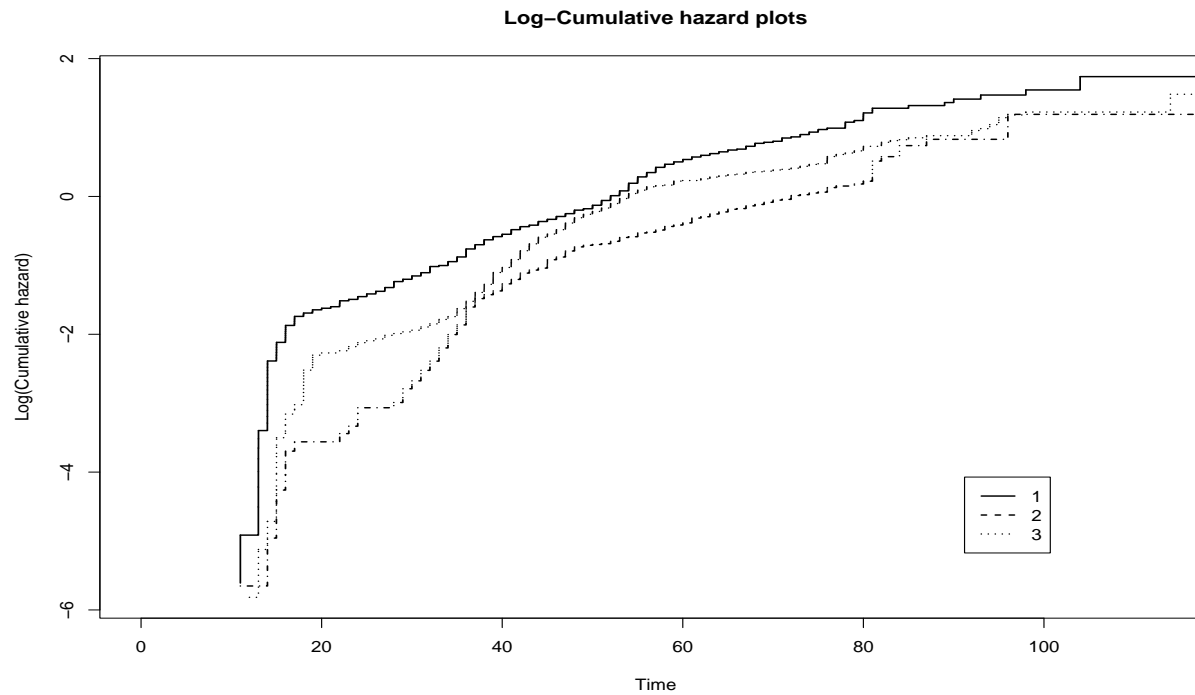


Seems to be ok.

# Checking Proportionality

How about the age groups:

```
> fit<-coxph(Surv(time,status==1)~strata(age.gr))  
> plotkum(fit)
```



The assumption about proportional effects seems to be violated.

# Checking Proportionality: tests

- Graphical test will reveal serious violations, but they may be difficult to use in general.
- If effect is expected to wear off with time or to increase with time, one may try to describe this effect and then fit a Cox model with a *time-varying* explanatory variable  $f(t)$

$$\exp(\beta + \theta \cdot f(t)) = \frac{\lambda_1(t)}{\lambda_2(t)}$$

For a given choice of  $f(t)$  we can then test if  $\theta$  is 0, i.e, if the time-varying effect can be left out of the model. A common choice of  $f(t)$  is

$$f(t) = \begin{cases} \log(t) & \text{IF GROUP}=1 \\ 0 & \text{IF GROUP}=0 \end{cases}$$

```
> fit<-coxph(Surv(time,status==1)~treatF+treatI)
> cox.zph(fit,trans='log')
      rho chisq      p
treatF 0.0191 0.295 0.5871
treatI 0.0607 2.921 0.0874
GLOBAL   NA 3.056 0.2170
```

So the departure from proportionality does not appear to be very pronounced.

# Proportionality In The General Cox-Model

The general Cox-model

$$\lambda_i(t) = \lambda_0(t) \exp(\beta_1 X_{i1} + \dots + \beta_p X_{ip})$$

In this model we wish to investigate if each of the covariates are consistent with the proportional hazards assumption. We stratify based on a grouping ( $k=1, \dots, K$ ) based on  $X_{i1}$ 's values and get a stratified Cox-model

$$\lambda_i(t) = \lambda_{0k}(t) \exp(\beta_2 X_{i2} + \dots + \beta_p X_{ip}) \text{ if } X_{i1} \in A_k \text{ } k\text{th stratum}$$

Now, if the underlying full Cox-model is true the baseline estimates  $h_{0k}(t)$  should be proportional, as

$$\lambda_{0k}(t) = \lambda_0(t) \exp\left(\sum_{k=1}^K \beta_{1k} I(X_{i1} \in A_k)\right)$$

where we multiply a proportionality depending on which group  $X_{i1}$  belongs to.

We therefore can make a graphical model-check of proportionality by making graphs of  $\log(\int_0^t \lambda_{0k}(s) ds)$ .

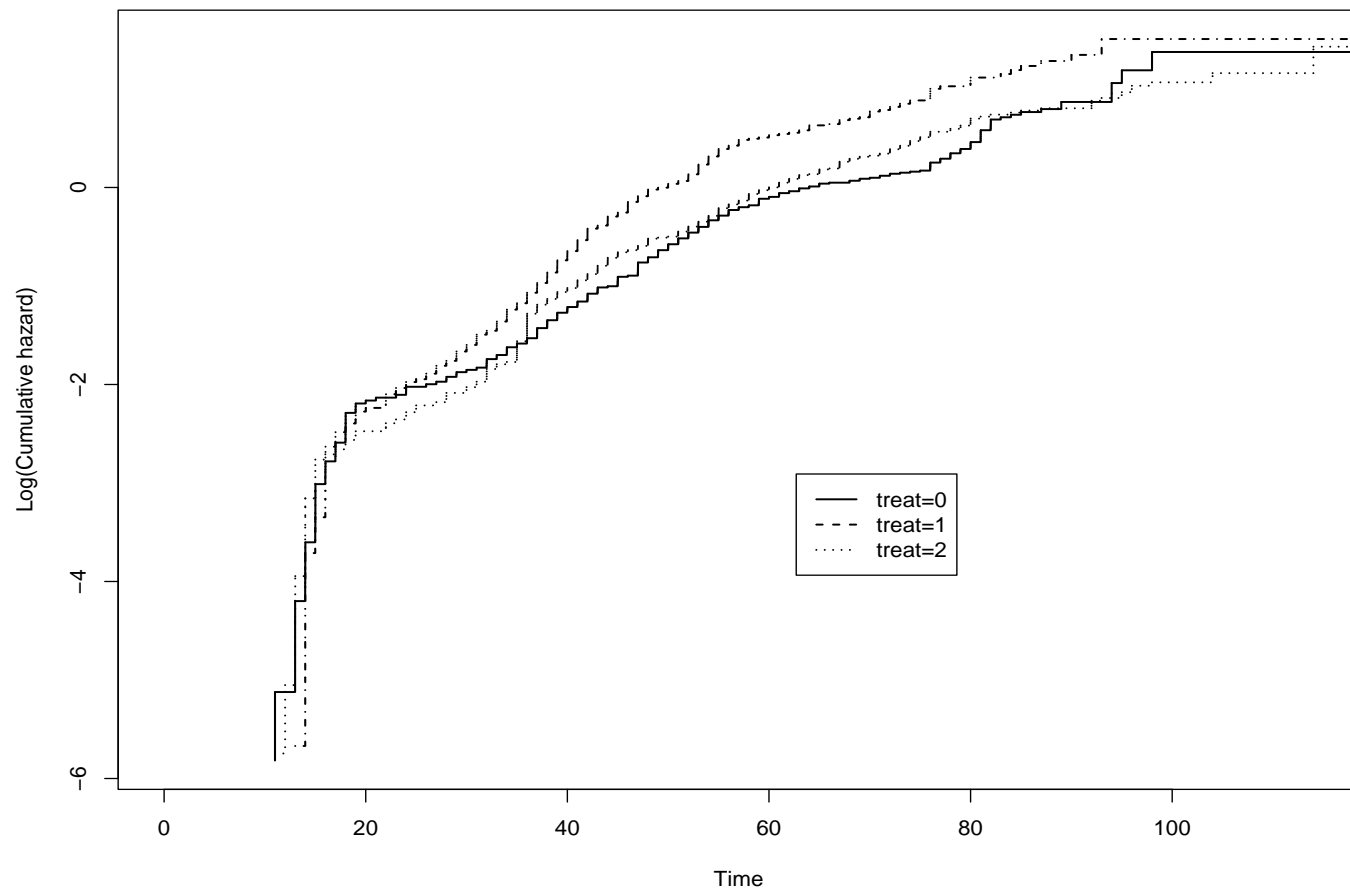
# Proportionality In The General Cox-Model

Considering the worm-data and the Cox-model with treatment and age.gr:

```
> fit<-coxph(Surv(time,status==1)~strata(treat)+factor(age.gr))  
> plotkum(fit)  
> fit<-coxph(Surv(time,status==1)~factor(treat)+factor(age.gr))  
> cox.zph(fit,transform='log')
```

	rho	chisq	p
factor(treat)1	0.0137	0.153	0.6958
factor(treat)2	0.0528	2.193	0.1386
factor(age.gr)1	0.0644	3.255	0.0712
factor(age.gr)2	0.0734	4.305	0.0380
GLOBAL	NA	7.765	0.1006

Log-Cumulative hazard plots



# Checking Assumptions: Summary

The Cox regression model is a very useful regression technique for survival data.

Remember that the conclusions based on the analysis only are valid when the model provides a reasonable description of the data. The model should therefore be validated.

- **Linearity** should be validated.
- Possible **interactions** should be investigated.
- The **proportionality** of the effects should be graphically assessed. When in doubt various test can be carried out.

# Exercise using R

- (1) Redo the analysis on slide 4-5. Compute the relative risks  $RR_{F:K}$ , comparing Fenb. to control, in the three age groups.

In the following we wish to use a model based on `treatment` and `kvart`. The latter gives the time where the cocoon is produced with values 0: dec-feb; 1: march-may; 2: june-august; 3: sept-nov.

- (2) Check the proportional hazards assumption for the Cox-model in this situation.
- (3) Propose a suitable model based on `treatment` and `kvart`, and analyse the data based on this model. Give relevant relative risk estimates associated with 95% confidence intervals.