

Survival analysis, Poisson regression and Cox regression I

Torben Martinussen

22. maj 2009

Modelling Survival Data

- ▶ To model and analyze survival data and deal with censorings in a convenient way one often specifies the model through the mortality rate (the intensity of death, the hazard rate):

$\lambda(t)$ = risk of dying among people at risk at time t

- ▶ The intensity is closely related to the survival probability

$$P(T > t) = S(t) = \exp\left(-\int_0^t \lambda(s)ds\right)$$

- ▶ in other words

$$-\log S(t) = \Lambda(t) = \int_0^t \lambda(s)ds$$

Parametric Models

- ▶ To give precise and clear answers. Kaplan-Meier curves and log-rank tests are not always sufficient.
- ▶ By making simplifying reasonable assumptions something more can be said through parametric models.
- ▶ The simplest possible survival model simply claims that the intensity is constant over time

$$\lambda(t) = \lambda, \quad \Lambda(t) = \lambda t$$

When the rate is constant the survival time is exponentially distributed.

- ▶ One may validate the model by considering the non-parametric estimator of the cumulative intensity (Nelson-Aalen)

$$\hat{\Lambda}(t)$$

- ▶ Should be approximately linear if the model is correct
- ▶ When this appears reasonable from the plot, we proceed to estimate λ and its standard error.

Estimation

- ▶ The maximum likelihood method gives that

$$\hat{\lambda} = \frac{D}{T}$$

where

D = #of occurrences

T = #total time at risk, exposure time

i.e. the occurrence-exposure ratio.

- ▶ It is better to construct a confidence interval on log-scale and transform it back to the original scale. It turns out that

$$se(\ln(\hat{\lambda})) = \frac{1}{\sqrt{D}}$$

- ▶ Transforming this back to the original scale we get

$$\left[\hat{\lambda} \exp\left(-\frac{1.96}{\sqrt{D}}\right); \hat{\lambda} \exp\left(+\frac{1.96}{\sqrt{D}}\right) \right]$$

The Exponential Survival Distribution

- ▶ Based on our estimate of λ we now have a guess on the survival function

$$\hat{S}(t) = \exp(-t\hat{\lambda})$$

and the expected survival time

$$\frac{1}{\hat{\lambda}}$$

- ▶ For the group as a whole we can analyse the data using **Poisson regression**. We start the analysis of constant intensities by considering how to compare constant intensities for K groups.
- ▶ Below, we shall see how the constant intensity model can be used more generally to describe mortality with piecewise constant rates.
- ▶ The simple constant rate model is appealing in that it is sufficient to keep track of the number of deaths and the total risk time.

Comparison of constant intensities

- ▶ Given K samples from different populations one may now wish to compare the intensities $(\lambda_1, \dots, \lambda_K)$. We estimate for each group separately

$$\hat{\lambda}_k = \frac{D_k}{T_k}$$

by occurrence-exposure rates and

- ▶ If groups all had the same mortality a combined estimate would be

$$\hat{\lambda} = \frac{\sum_k D_k}{\sum_k T_k}$$

the combined occurrence-exposure rate.

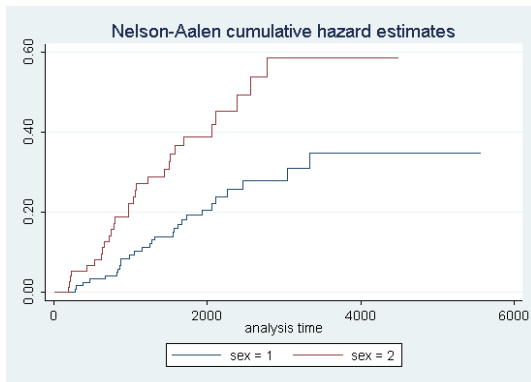
- ▶ A test for $H_0 : \lambda_1 = \dots = \lambda_K$ is given by

$$2 \sum_k D_k (\ln(\hat{\lambda}_k) - \ln(\hat{\lambda}))$$

Which is approximately χ^2 with $K - 1$ degrees of freedom (get back to this).

Melanoma data

- Consider the Nelson-Aalen plot for the groups by sex:



- We have: $\hat{\lambda}_M = 0.069 \text{ years}^{-1}$ and $\hat{\lambda}_F = 0.036 \text{ years}^{-1}$, and hence an estimated relative risk of $\hat{\lambda}_M / \hat{\lambda}_F = 1.94$.
- Test of $H_0 : \lambda_M = \lambda_F$ gives $p = 0.01$.

Example in Stata

Using software, Poisson-regression; Estimates $\log(\lambda)$.

```
xi: poisson dead i.sex, exposure(days)
i.sex          _Isex_1-2          (naturally coded; _Isex_1 omitted)
```

dead	Coef.	Std. Err.	z	P> z	[95% Conf. Interval]	
_Isex_2	.6616329	.2649472	2.50	0.013	.1423459	1.18092
_cons	-9.237167	.1889822	-48.88	0.000	-9.607565	-8.866768
days	(exposure)					

```
. xi: poisson dead i.sex, exposure(days) irr
```

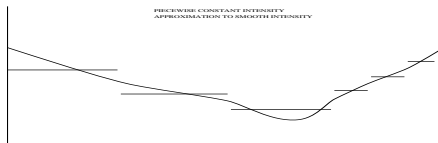
dead	IRR	Std. Err.	z	P> z	[95% Conf. Interval]	
<hr/>						
_Isex_2	1.937954	.5134557	2.50	0.013	1.152975	3.25737
days	(exposure)					

```
. display exp(0.66163-9.237167)*365
.06886277
```

```
. display exp(-9.237167)*365
.03553385
```


Piecewise constant hazard rates

- ▶ The simple constant hazard rate model may be extended by allowing piecewise constant but different hazard rates. This may provide a sensible summary of many phenomena.



- ▶ Such that

$$\lambda(t) = \lambda_k \text{ for } t \in [c_{k-1}, c_k]$$

- ▶ We only need to keep track of the total number deaths and the exposure time in each time interval.
- ▶ Estimates and standard errors are given as before, e.g.,

$$\hat{\lambda}_k = \frac{D_k}{T_k} = \frac{\text{no of occur. in } [c_{k-1}, c_k]}{\text{exposure time in } [c_{k-1}, c_k]}$$

Example: Smoking and mortality

- ▶ Cohort study of smoking and mortality:

Age-group	Dead	Person-ys	Smoker
35-44	32	52407	1
35-44	2	18790	0
45-54	104	43248	1
45-54	12	10673	0
54-64	206	28612	1
54-64	28	5712	0
65-74	186	12663	1
65-74	28	2585	0
75-84	102	5317	1
75-84	31	1462	0

- ▶ Disregard for a moment the grouping given by age and consider only influence of smoking.

Smoker	Deaths	Person-ys (exposure)	$\lambda \cdot 1000$	$se(\lambda) \cdot 1000$
1	630	142247	4.43	0.18
0	101	39222	2.58	0.26

Example: Smoking and mortality

► $\hat{\lambda} \cdot 1000 = (1000 \cdot 731)/181469 = 4.03.$

► Stata

```
iri 630 101 142247 39222
```

	Exposed	Unexposed	Total	
Cases	630	101	731	
Person-time	142247	39222	181469	
Incidence Rate	.0044289	.0025751	.0040282	
	Point estimate		[95% Conf. Interval]	
Inc. rate diff.	.0018538		.0012441	.0024636
Inc. rate ratio	1.71991		1.392063	2.143639 (exact)

Example: Smoking and mortality

Using software

```
xi: poisson dead i.smoke, exposure(followup)
i.smoke      _Ismoke_0-1      (naturally coded; _Ismoke_0 omitted)
```

dead	Coef.	Std. Err.	z	P> z	[95% Conf. Interval]	
<hr/>						
_Ismoke_1	.5422721	.1071834	5.06	0.000	.3321964	.7523478
_cons	-5.961873	.0995037	-59.92	0.000	-6.156896	-5.766849
followup	(exposure)					

```
display exp(0.5422721-5.961873)
.00442891
. display exp(-5.961873)
.00257508
. * Compare with iri-output!
. display exp(0.5422721)
1.7199102
. xi: poisson dead i.smoke, exposure(followup) irr
```

dead	IRR	Std. Err.	z	P> z	[95% Conf. Interval]	
<hr/>						
_Ismoke_1	1.71991	.1843459	5.06	0.000	1.394027	2.121976
followup	(exposure)					

Relative risk: $\exp(0.5423) = 1.72$.
Effect of smoking?

Poisson Model

- ▶ The basis for the *Poisson* model is the *incidence rate* (intensity) denoted λ , which is the expected amount of events per time unit.
- ▶ How the intensity depends on various covariates can be analysed by Poisson regression.
- ▶ The log-linear regression model for Poisson counts models the incidence rate

$$\log(\lambda) = \beta_0 + \beta_1 X_{i1} + \dots + \beta_p X_{ip}$$

for covariates X_{i1}, \dots, X_{ip} (explanatory variables), and with β_0 the baseline level.

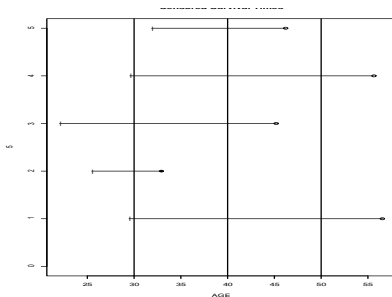
- ▶ β_1, \dots, β_p are the regression-coefficients, that represent the effects of the covariates.
- ▶ β_1 is the effect of X_{i1} when we have corrected for the other covariates (when these are fixed).

Poisson Models and Mortality

- ▶ Poisson models may be used to study mortality while recognizing the time aspect.
- ▶ If we believe that the death rate is constant in age groups (20-30, 30-40,...,say, which is approximately true) then the total number of deaths in the various age groups are poisson with expected number of deaths in age group j :

$$\lambda_j \cdot T_j$$

where T_j is the total exposure time for age group j .



Poisson Models

- ▶ Cohort study of smoking and mortality. Easy to correct for age-differences. Interpret below output!

```
xi: poisson dead i.smoke i.age, exposure(followup)
i.smoke      _Ismoke_0-1      (naturally coded; _Ismoke_0 omitted)
i.age        _Iage_40-80      (naturally coded; _Iage_40 omitted)

-----+-----
             dead |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-----+-----
    _Ismoke_1 |   .3546374   .1073739     3.30   0.001   .1441884   .5650864
    _Iage_50  |   1.484002   .1951034     7.61   0.000   1.101606   1.866397
    _Iage_60  |   2.627454   .1837271    14.30   0.000   2.267356   2.987553
    _Iage_70  |   3.350485   .1847992    18.13   0.000   2.988285   3.712685
    _Iage_80  |   3.700092   .1922195    19.25   0.000   3.323349   4.076836
    _cons     |  -7.919407   .1917625   -41.30   0.000   -8.295255  -7.543559
followup     | (exposure)
```

- ▶ We start by considering the fit of the model.
- ▶ Wish to calculate the expected number of deaths of non-smoking 80 years under the model.

Poisson Models

- ▶ To get estimate of $\log(\lambda)$:

```
lincom _cons + _Iage_80
```

```
( 1)  [dead]_Iage_80 + [dead]_cons = 0
```

dead	Coef.	Std. Err.	z	P> z	[95% Conf. Interval]	
-----+-----						
(1)	-4.219315	.1249846	-33.76	0.000	-4.46428	-3.974349

- ▶ To get estimate of expected number of deaths and 95%-ci:

```
. display exp(-4.219315)*1462
21.504144
. display exp(-4.46428)*1462
16.83198
. display exp(-3.974349)*1462
27.473218
```

- ▶ Note that group of non-smoking 80 year olds die quite a bit more than expected by the model. 31 is outside the confidence interval for the count (under the model) [16.8, 27.5].

Likelihood ratio test

- ▶ A general strategy for testing is based on the **likelihood principle**.
- ▶ A likelihood is a number that reflects how well the model fits the data
- ▶ Different models may be compared by comparing their corresponding likelihoods.
- ▶ The models needs to be nested in the sense that one model contains the other as a special case.
- ▶ The likelihood ratio test is

$$-2\ln(Q) = -2(l_2 - l_1)$$

which is approximately χ^2 with DF degrees. DF is the number of parameters that can be omitted in the simpler model.

► Check for interaction

```
. quietly xi: poisson dead smoke i.age, exposure(followup)
. estimates store modell
. quietly xi: poisson dead i.smoke i.age i.smoke*i.age, exposure(followup)
. estimates store model2
. lrtest modell model2
Likelihood-ratio test                                LR chi2(4) =      12.13
(Assumption: modell nested in model2)                Prob > chi2 =      0.0164
```

```
. xi: poisson dead i.smoke i.age i.smoke*i.age, exposure(followup)
```

dead	Coef.	Std. Err.	z	P> z	[95% Conf. Interval]
__Ismoke_1	1.746873	.7288689	2.40	0.017	.3183163 3.17543
__Iage_50	2.357367	.7637625	3.09	0.002	.8604198 3.854314
__Iage_60	3.829813	.731925	5.23	0.000	2.395266 5.264359
__Iage_70	4.622656	.731925	6.32	0.000	3.18811 6.057203
__Iage_80	5.294359	.7295601	7.26	0.000	3.864448 6.724271
_IsmoXage~50	-.9866227	.7900624	-1.25	0.212	-2.535117 .5618712
_IsmoXage~60	-1.362458	.7561868	-1.80	0.072	-2.844557 .1196405
_IsmoXage~70	-1.44229	.7565319	-1.91	0.057	-2.925065 .0404855
_IsmoXage~80	-1.846991	.7571736	-2.44	0.015	-3.331024 -.3629584
__cons	-9.147933	.7071067	-12.94	0.000	-10.53384 -7.762029
followup	(exposure)				

Poisson Models

► Estimates:

*Non-smokers (If 7.40106 is added to all then we get fig. in handouts
* so that rate for smoker in first age band is set to 0).

```
lincom _cons  
lincom _cons+_Iage_50  
lincom _cons+_Iage_60  
lincom _cons+_Iage_70  
lincom _cons+_Iage_80
```

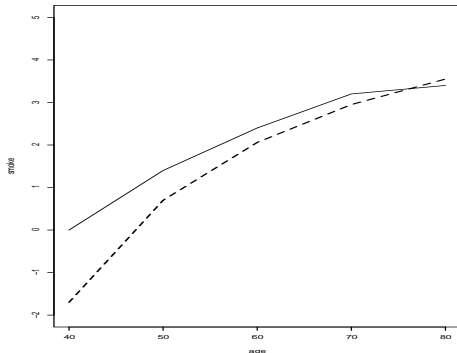
*Smokers

```
lincom _cons+ _Ismoke_1  
lincom _cons+ _Ismoke_1+_Iage_50+ _IsmoXage_1_50  
lincom _cons+ _Ismoke_1+_Iage_60+ _IsmoXage_1_60  
lincom _cons+ _Ismoke_1+_Iage_70+ _IsmoXage_1_70  
lincom _cons+ _Ismoke_1+_Iage_80+ _IsmoXage_1_80
```

	dead	Coef.	Std. Err.	z	P> z	[95% Conf. Interval]
(1)	-9.147933	.7071067	-12.94	0.000	-10.53384	-7.762029
(1)	-6.790566	.2886751	-23.52	0.000	-7.356359	-6.224773
(1)	-5.31812	.1889822	-28.14	0.000	-5.688518	-4.947722
(1)	-4.525276	.1889822	-23.95	0.000	-4.895675	-4.154878
(1)	-3.853573	.1796053	-21.46	0.000	-4.205593	-3.501554
(1)	-7.40106	.1767767	-41.87	0.000	-7.747536	-7.054584
(1)	-6.030315	.0980581	-61.50	0.000	-6.222506	-5.838125
(1)	-4.933705	.0696733	-70.81	0.000	-5.070262	-4.797148
(1)	-4.220693	.0733236	-57.56	0.000	-4.364404	-4.076981
(1)	-3.953692	.0990148	-39.93	0.000	-4.147757	-3.759626

Example Continued

A graph showing the log incidence rates and the interaction, reveal what is going on:



So the difference between smokers and non-smokers decrease with age (hence the interaction). The difference is particular large for 40 year olds.

Cox-Regression

- ▶ Cox-Regression is the regression technique for survival analysis ;
- ▶ Regression techniques are very useful for dealing with many covariates;
- ▶ Can be used to learn about treatment effect while correcting for other covariates.

Hazard function

$$\lambda_i(t)dt = P(T_i \in [t, t + dt] \mid \text{alive at time } t)$$

Cox model:

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

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

Note: $\lambda_0(t)$ is not further specified!

The Cox model

- ▶ The regression coefficients β_1, \dots, β_p represent the effects of the covariates.
- ▶ β_1 is the effect of X_{i1} when we have corrected for the other covariates.
- ▶ β_1 may be interpreted in terms of the **relative risk** when the covariate X_{i1} is increased 1:

$$\frac{\lambda_0(t) \exp(\beta_1(X_{i1} + 1) + \dots + \beta_p X_{ip})}{\lambda_0(t) \exp(\beta_1 X_{i1} + \dots + \beta_p X_{ip})} = \exp(\beta_1)$$

- ▶ If $\beta_1 > 0$ the risk of dying increases as X_{i1} increases, and if $\beta_1 < 0$ the risk of dying decreases as X_{i1} increases.
- ▶ The quantity

$$\hat{\beta}_1 X_{i1} + \dots + \hat{\beta}_p X_{ip}$$

is called the prognostic index for the i th subject.

Two-Sample Cox Model

- ▶ Consider a simple 2-sample situation where we wish to study effect of sex.
- ▶ Defining a covariate for the i th patient

$$X_i = \begin{cases} 1 & \text{Male} \\ 0 & \text{Female} \end{cases}$$

- ▶ The hazard can be written as

$$\lambda_i(t) = \lambda_0(t) \exp(\beta X_i)$$

giving the contributions $\lambda_0(t) \exp(\beta)$ when $X_i = 1$ and $\lambda_0(t)$ when $X_i = 0$.

- ▶ Note again that $\lambda_0(t)$ is unspecified meaning that which distribution is unspecified?

Two-Sample Cox Model

The Cox model is fitted to the data as follows

```
stset days, failure(status==1)
      xi: stcox i.sex
i.sex      _Isex_1-2      (naturally coded; _Isex_1 omitted)

Cox regression -- no ties

No. of subjects =          205      Number of obs   =          205
No. of failures =           57
Time at risk    =        441324

Log likelihood   =   -280.12397      LR chi2(1)      =          6.15
                                      Prob > chi2      =          0.0131
```

	_t	Haz. Ratio	Std. Err.	z	P> z	[95% Conf. Interval]	
	_Isex_2	1.939011	.5140979	2.50	0.013	1.153182	3.260339

- ▶ The Wald test is given as $\left(\hat{\beta}/SE(\hat{\beta})\right)^2$ which is χ^2_1 .
- ▶ How should the value of 1.93 be interpreted?

Example: Melanoma data

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

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

We get:

```
. xi: stcox i.sex lthick
i.sex          _Isex_1-2          (naturally coded; _Isex_1 omitted)

      failure _d:  status == 1
      analysis time _t:  days

No. of subjects =          205          Number of obs   =          205
No. of failures =           57
Time at risk    =        441324

Log likelihood   =    -266.4747          LR chi2(2)      =        33.45
                                          Prob > chi2      =        0.0000
```

	_t	Haz. Ratio	Std. Err.	z	P> z	[95% Conf. Interval]
	+					
	_Isex_2	1.580893	.4247382	1.70	0.088	.9337066 2.676669
	lthick	2.183408	.3435457	4.96	0.000	1.603997 2.972119

What does the relative risks of 1.58 and 2.18 mean?

Survival predictions based on Cox-model

- ▶ Let X^0 denote the covariates for a given subject.
- ▶ The survival function assuming the Cox-model is given by

$$P(T^0 > t) = \exp\left(-\int_0^t \lambda(s, X^0) ds\right) = \exp(-\Lambda_0(t)e^{PI(\beta)}),$$

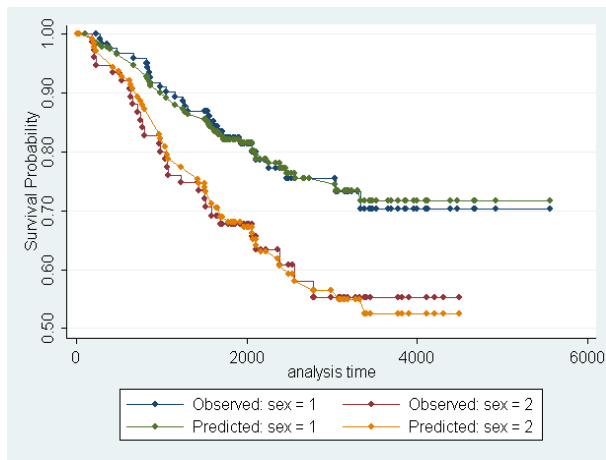
where $PI(\beta) = \beta_1 X_1^0 + \dots + \beta_p X_p^0$ is called the prognostic index.

- ▶ The survival function may thus be estimated by insertion of $\hat{\beta}$ and $\hat{\Lambda}_0(t)$.

Example: Melanoma data

- Only sex in model, and compare with Kaplan-Meier estimate.

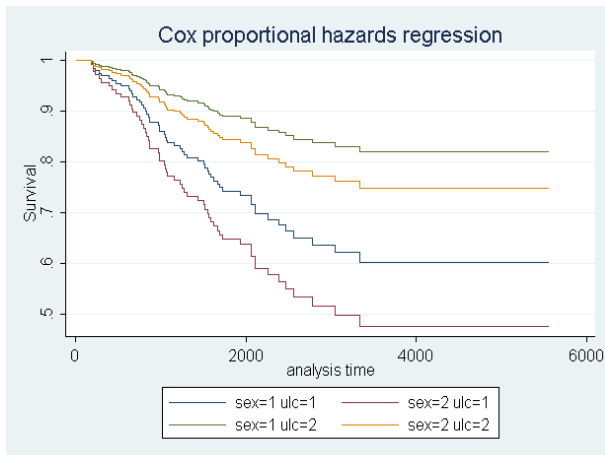
```
. stcoxkm, by (sex)
```



Example: Melanoma data

- Several variables in the model (what is the value of lthick?):

```
. stcox sex ulc lthick, basesurv(s)  
. stcurve, survival at1(sex=1 ulc=1) at2(sex=2 ulc=1)/*  
  */ at3(sex=1 ulc=2) at4(sex=2 ulc=2)
```



Example: Melanoma data

► Several variables in the model:

```
. stcurve, survival at1(sex=1 ulc=2 lthick=4) at2(sex=2 ulc=2 lthick=4)/*  
> */ at3(sex=1 ulc=2 lthick=6) at4(sex=2 ulc=2 lthick=6)
```

