Survival analysis II

Research Methods for Quantitative Data 2018

Camille Lassale
Department of Epidemiology and Public Health

c.lassale@ucl.ac.uk

Objectives

By the end of this session you should be able to:

- Understand the concept of a hazard
- Fit a Cox Proportional Hazards model, interpret results and explain the meaning of a hazard ratio
- Describe methods to check the assumption of proportional hazards in the Cox model

Regression - survival analysis

- We saw previously how to estimate the proportion experiencing an event using Kaplan Meier methods, and how to compare responses between two groups using the log rank test

- We may wish to extend to regression models if
  - several explanatory variables
  - some continuous explanatory variables

- Most commonly used regression model in Epidemiology for time to event data is Cox Proportional Hazards Regression

Cox proportional hazards models

•Technique for modelling survival data

•Takes account of censoring

Enables us to

- **describe** the association of different prognostic factors (covariates) with survival
- **adjust** the effect of one prognostic factor for the effect of another
- **predict** individuals with better or worse prognosis

Cox proportional hazards models

- Produces estimates of hazard ratios for explanatory variables (sometimes known as rate ratios or risk ratios)
- Regression coefficients from the Cox model are on log scale (similarly to logistic regression)
- Exponentiate coefficient to obtain a hazard ratio
- *What is a hazard ratio? First define the hazard*

The hazard rate 1

- We saw in Survival Analysis 1 that an incidence rate represents the propensity to experience an event over the entire follow-up period

- It assumes that the rate of events remains constant

- However, the propensity to develop the event may not be constant over follow-up (as is assumed from the way we calculated the rate) – e.g. death rates increase over the age of 40

The hazard rate 2

- The **hazard rate** at a time point is the propensity to develop the event *at that instant in time* given the event not yet occurred

- It represents the instantaneous failure rate at time t

- It is also called **hazard function λ(t)**
- The shape of λ(t) will depend on the group under study.

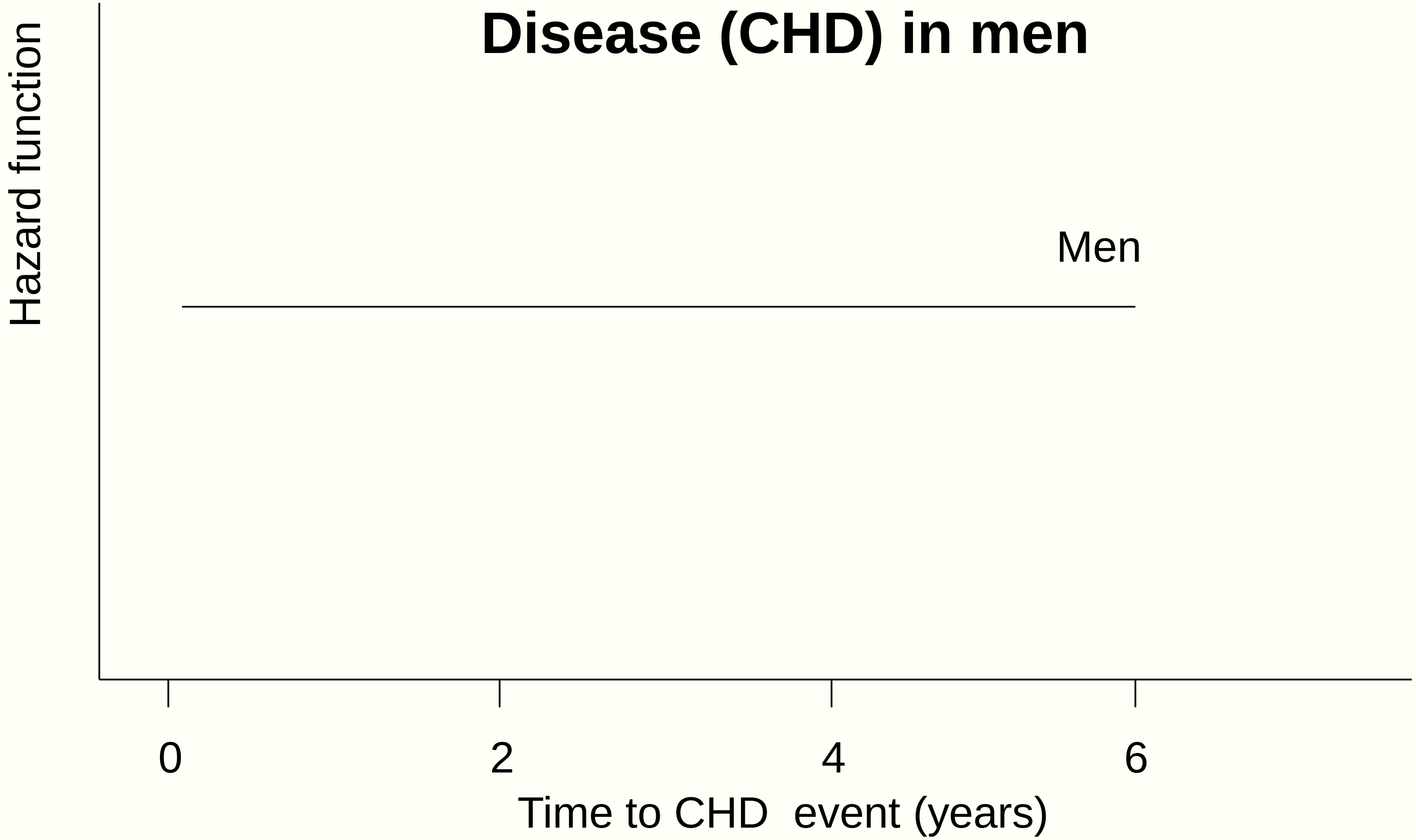
The hazard rate – Example

Week	1	2	3	4	5	6	7	8	9	Over all
No. people at risk	1000	999	994	986	976	964	952	946	943	1000
Person-weeks	1000	999	994	986	976	964	952	946	943	8760
No. with event in week	1	5	8	10	12	12	6	3	1	58
Event hazard rate	0.001	0.005	0.008	0.010	0.012	0.013	0.006	0.003	0.001	

Overall event rate = 58/8760 = 0.066 /person-week

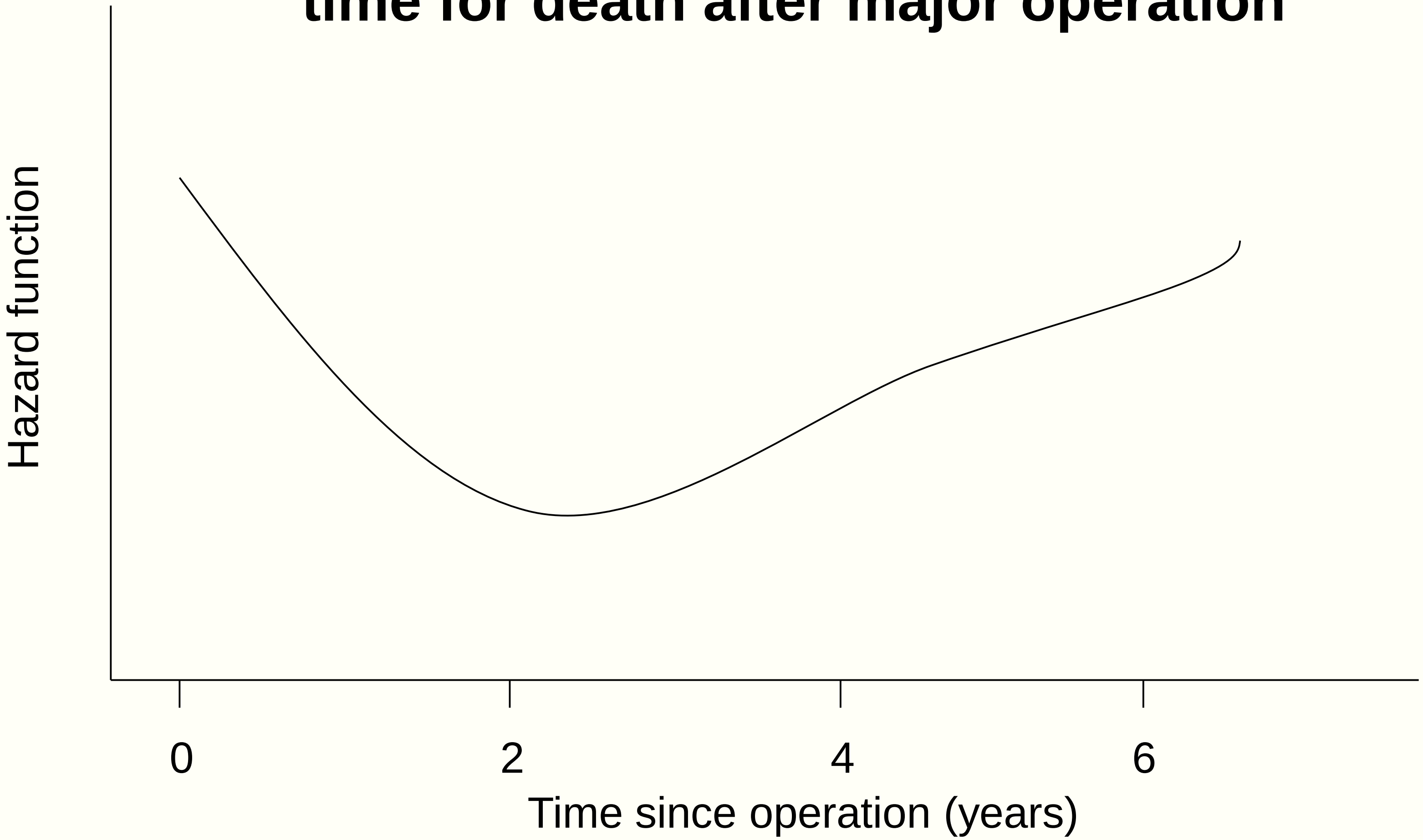
Hazard function – Example

Hypothetical constant hazard function over time for incidence of Coronary Heart Disease (CHD) in men



Hazard function – Example

Hypothetical hazard function varying over time for death after major operation



The hazard ratio

- The hazard rate in two groups can be compared using the **hazard ratio (HR)**
- It can be interpreted similarly to the odds ratio, risk ratio and rate ratio
- It is only reasonable to calculate the HR if the effect of each covariate on the outcome is assumed to be the same at all time points, irrespective of the underlying hazard at each time point – the **proportional hazards assumption**

Hazard ratio – Example

Week	1	2	3	4	5	6	7	8	9
Hazard rate group 1	0.002	0.006	0.012	0.016	0.030	0.026	0.015	0.007	0.001
Hazard rate group 2	0.001	0.005	0.008	0.010	0.012	0.013	0.006	0.003	0.001
Time-specific hazard ratio	2.0	1.6	1.5	1.6	2.5	2.0	2.5	2.3	1.0

Proportional hazards – assume that the true HR is the same in each period

The HR is then an average of the time-specific HRs – in this case HR ≈ 2

Hazard ratio – Example

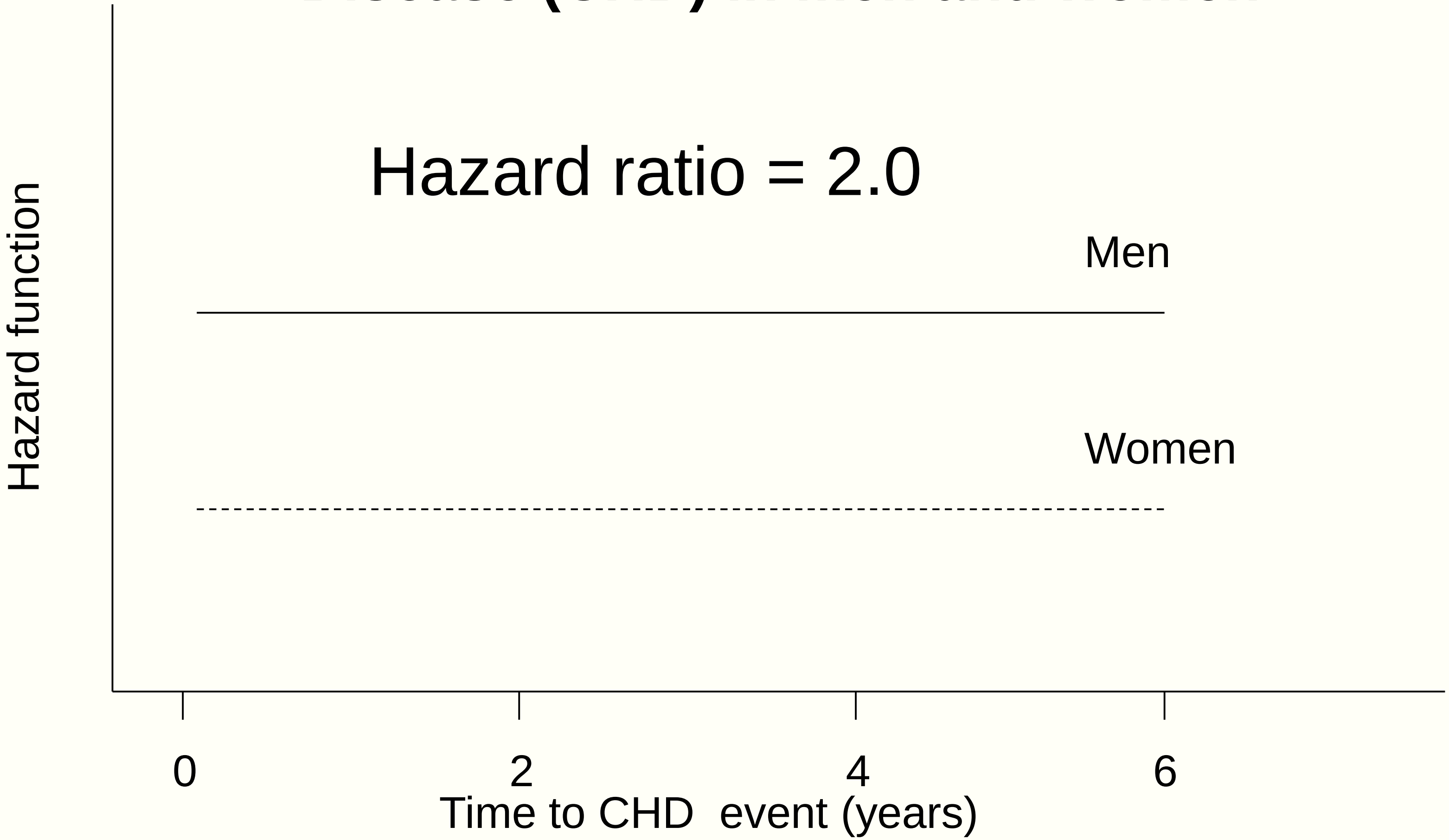
Week	1	2	3	4	5	6	7	8	9
Hazard rate group 1	0.002	0.006	0.012	0.016	0.030	0.045	0.065	0.077	0.093
Hazard rate group 2	0.001	0.003	0.005	0.007	0.010	0.014	0.020	0.025	0.030
Time-specific hazard ratio	2.00	2.00	2.40	2.29	3.00	3.21	3.25	3.08	3.10

The time-specific hazard is not constant. The effect of being in group 1 increases over time

The proportional hazards assumption does not hold, and it is not reasonable to calculate a HR for the entire time period

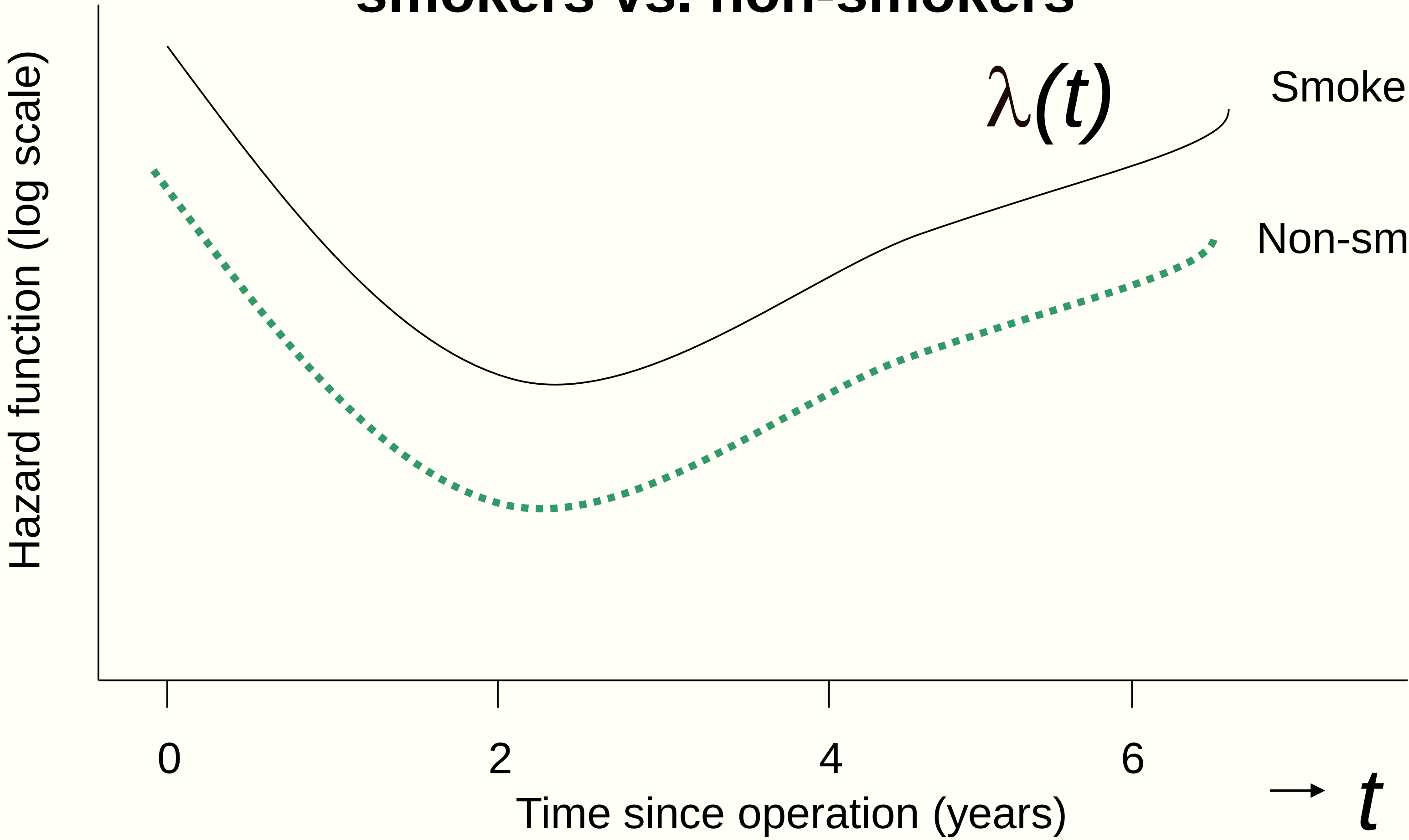
Hazard ratio – Example

Hypothetical constant hazard function over time for incidence of Coronary Heart Disease (CHD) in men and women



Hazard ratio – Example

Hypothetical hazard function varying over time for death after major operation: smokers vs. non-smokers



Cox regression

- The Cox model does not make any underlying assumptions about the shape of the underlying hazard function

- It estimates the underlying hazard function from the data

- It assumes proportional hazards for covariates (explanatory variables)

Smoking and CHD

	Hazard ratio (95% CI)	*p-value*
Non-smoker	**1**	
Smokes		
> 20 a day	**1.69** (1.03, 2.76)	*0.038*

Cox regression

- Assumes proportional hazards

- What does this mean?
  - In smoking and CHD example, the estimated hazard ratio for heavy smokers versus non-smokers is 1.69
  - This hazard ratio is assumed to be constant over time
  - At any time point, the hazard of an event for heavy smokers is 1.69 times the hazard for a non-smoker
  - The risk of CHD event for smokers is increased by 69% (compared to non-smokers)

Cox regression model

- $\lambda\ _i(t) = \lambda\ _0(t)exp(\beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_n \mathrm{xin})$
- where
  - $\lambda_i(t)$ hazard function for individual *i*
  - $\lambda_0(t)$ baseline hazard function

  - $x_1 \ldots x_n$ are covariates
  - $\beta_1 \ldots \beta_n$ are estimated effects of covariates (assumed constant over time)

Cox regression model

- $\lambda\ _i(t) = \lambda\ _0(t) exp(\beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_n xin)$

- where
  - $\lambda_i(t)$ hazard function for individual *i*
  - $\lambda_0(t)$ baseline hazard function CAN TAKE ANY FORM AND ESTIMATED FROM DATA (NON-PARAMETRIC)
  - $x_1 \ldots x_n$ are covariates
  - $\beta_1 \ldots \beta_n$ are estimated effects of covariates (assumed constant over time) PROPORTIONAL HAZARDS ASSUMPTION (PARAMETRIC)

Cox regression in Stata

Social position (low, middle and high grade) and association with all cause mortality

xi: stcox i.agegrp i.sex i.empgrade

```
•   No. of subjects =         10297                    Number of obs    =     10297
•   No. of failures =           605
•   Time at risk    =   176515.5127
•                                                      LR chi2(6)       =    225.94
•   Log likelihood  =    -5410.9155                    Prob > chi2      =    0.0000
•   ------------------------------------------------------------------------------
•             _t | Haz. Ratio   Std. Err.      z     P>|z|     [95% Conf. Interval]
•   -------------+----------------------------------------------------------------
•     _Iagegrp_2 |   1.294235   .1979559     1.69   0.092     .9590035    1.746651
•     _Iagegrp_3 |    2.29277   .3291687     5.78   0.000     1.730434    3.037847
•     _Iagegrp_4 |   4.110989   .5268814    11.03   0.000     3.197811    5.284937
•        _Isex_2 |   .6950344   .0706329    -3.58   0.000     .5695121    .8482223
•   _Iempgrade_2 |   1.308157   .1366984     2.57   0.010     1.065888     1.60549
•   _Iempgrade_3 |    1.97288   .2480098     5.41   0.000     1.542043     2.52409
•   ------------------------------------------------------------------------------
```

Hazard ratio (low vs. high employment grade)=1.97 (95% CI 1.54 to 2.52) p<0.001

Interpreting hazard ratio

- The hazard is the rate at which events occur

- Suppose you have two treatments (A and B)

Hazard ratio =

hazard (treatment A)/hazard(treatment B)

- **HR=1** (treatments are the same)
- **HR < 1** hazard rate smaller with treatment A
- **HR > 1** hazard rate smaller with treatment B

Interpreting hazard ratios

- Employees in low employment grades have an increased risk of death (at any time during follow up), compared to people in high employment grades, after controlling for age and sex (hazard ratio [HR] = 1.97)

- Employees in low employment grades have about twice the risk of premature death compared to high employment grades

Cox regression in Stata

Obtaining coefficients on log scale (nohr)

xi: stcox i.agegrp i.sex i.empgrade,nohr

```
•   No. of subjects =        10297                  Number of obs   =     10297
•   No. of failures =          605
•   Time at risk    =  176515.5127
•                                                   LR chi2(6)      =    225.94
•   Log likelihood  =   -5410.9155                  Prob > chi2     =    0.0000

•   ------------------------------------------------------------------------------
•             _t |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
•   -------------+----------------------------------------------------------------
•     _Iagegrp_2 |   .2579199    .152952     1.69   0.092    -.0418605    .5577004
•     _Iagegrp_3 |   .8297607   .1435681     5.78   0.000     .5483724    1.111149
•     _Iagegrp_4 |   1.413664   .1281642    11.03   0.000     1.162466    1.664861
•        _Isex_2 |  -.3637939    .101625    -3.58   0.000    -.5629753   -.1646125
•   _Iempgrade_2 |   .2686189    .104497     2.57   0.010     .0638086    .4734292
•   _Iempgrade_3 |   .6794945   .1257095     5.41   0.000     .4331084    .9258807
•   ------------------------------------------------------------------------------
```

Coefficient (low vs. high employment grade)=0.679
Hazard ratio =exp(0.679)=1.97

Comparing models

Using likelihood ratio test

Is employment grade related to all cause mortality after adjusting for baseline illness?

```
stcox i.agegrp i.sex i.empgrade i.ill
est store a
stcox i.agegrp i.sex i.ill
lrtest a
```

```
Likelihood-ratio test                 LR chi2(2)  =   25.61
(Assumption: . nested in a)           Prob > chi2 =   .0000
```

Fitting interaction terms

Fit and test for interaction between sex and employment grade

```
xi: stcox i.agegrp i.sex*i.empgrade
estimates store a
xi: stcox i.agegrp i.sex i.empgrade
lrtest a
```

Cox regression in Stata

Estimates of the baseline hazard function

```
xi: stcox i.empgrade, basesurv(s) basehc(h) basech(ch)
```

basesurv stores estimate of baseline survival function:

estimated probability of surviving till time *t for all covariates equal to 0*

basech stores estimate of cumulative hazard function

for all covariates equal to 0

Testing proportional hazards assumption

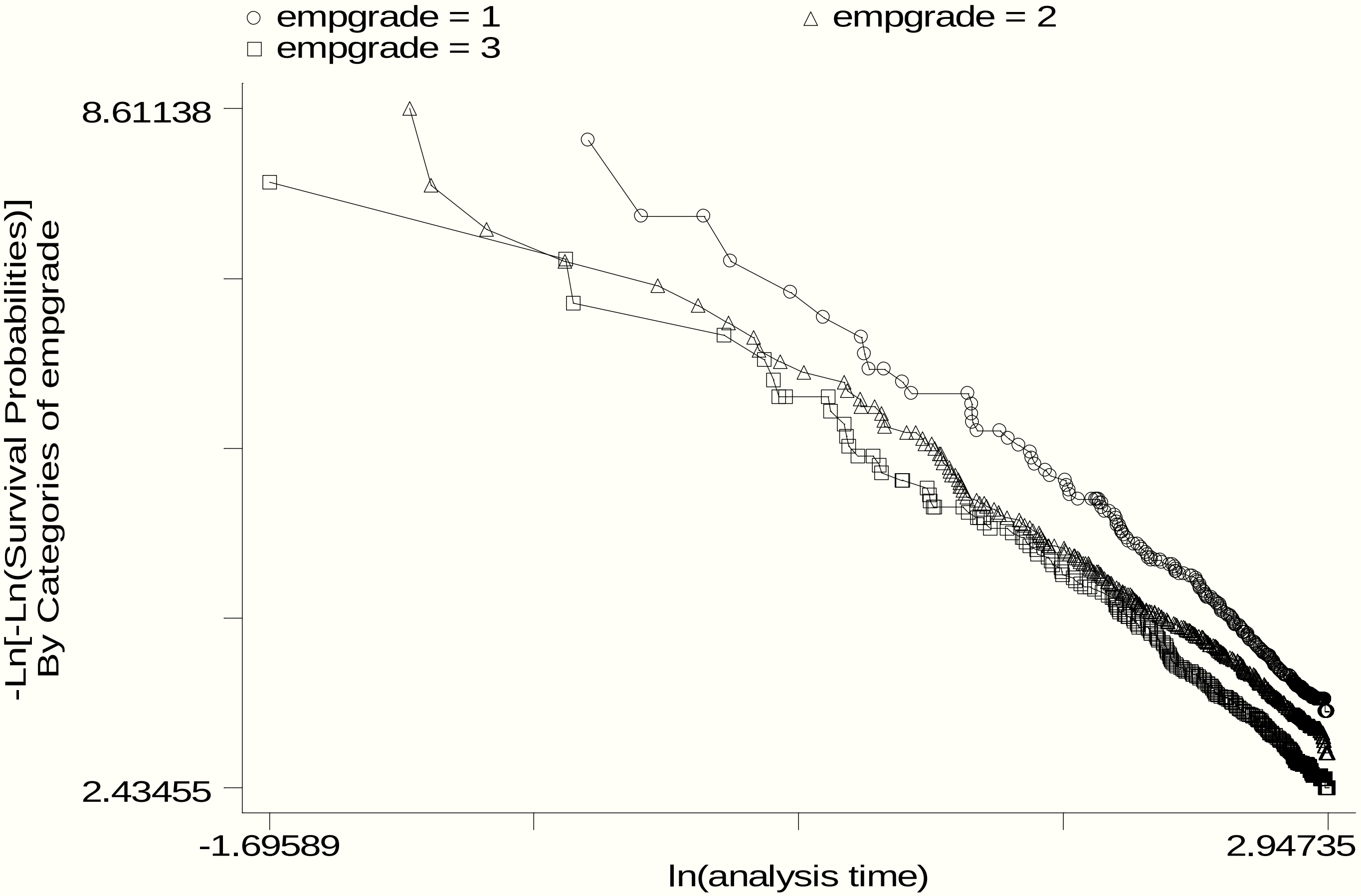
1. Graphical methods
   a. Comparison of Kaplan-Meier estimates by group

   b. Plot estimates **log cumulative baseline hazards** for each group against **time** (curves should be parallel)

   c. Plot **minus the log cumulative baseline hazard** for each group against **log survival time** (easier to judge if lines are parallel)

Testing proportional hazards assumption (1c)

stphplot, strata(empgrade) adjust(sex agegrp)



Testing proportional hazards assumption

2. Formal test of proportional hazards assumption

   a. Include an interaction between a covariate and a function of time

   Log(time) often used, but could be any function of analysis time

   b. Based on residuals

   e.g. Stata has a test based on a type of residual known as Schoenfeld residuals

Testing proportional hazards assumption (2a)

xi: stcox i.empgrade, tvc(i.empgrade) texp(log(_t))

```
------------------------------------------------------------------------------
          _t | Haz. Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
rh           |
             
_Iempgrade_2 |   2.367957   .8786467     2.32   0.020     1.144273    4.900247
_Iempgrade_3 |   2.577014   1.047479     2.33   0.020     1.161794    5.716164
-------------+----------------------------------------------------------------
t            |
_Iempgrade_2 |   .7678917   .1203219    -1.69   0.092     .5648383    1.043941
_Iempgrade_3 |   .8898324   .1507174    -0.69   0.491     .6384605    1.240173
------------------------------------------------------------------------------

Note: second equation contains variables that continuously vary with respectto
   time; variables are interacted with current values of log(_t).
```

```
estimates store d
lrtest d c
Likelihood-ratio test                                 LR chi2(2)  =      3.35
(Assumption: c nested in d)                           Prob > chi2 =    0.1873
```

Testing proportional hazards assumption (2b)

xi: stcox i.empgrade i.sex,schoenfeld(sch*) scaledsch(sca*)

estat phtest, log detail

```
Test of proportional-hazards assumption

Time:  Log(t)
----------------------------------------------------------------
            |       rho            chi2       df       Prob>chi2
------------+---------------------------------------------------
_Iempgrade_2|     -0.07845         3.64        1         0.0565
_Iempgrade_3|     -0.04932         1.51        1         0.2190
_Isex_2     |      0.05494         1.98        1         0.1596
------------+---------------------------------------------------
global test |                      5.33        3         0.1489
----------------------------------------------------------------
```

When proportional hazards assumption is not met

- STRATIFY the analysis (two options for categorical data)

A. Fit separate models for each stratum group
   - Both baseline hazard rate and hazard ratios vary by group
   - Will obtain a separate HR for each stratum group

e.g. fit separate Cox models for men and women

HR (smokers vs. non smokers): Men 2.0, Women 1.5

When proportional hazards assumption is not met

- STRATIFY the analysis (two options for categorical data)

B. Allow baseline hazards by group to vary, but assume covariate effects are same across strata. This allows underlying hazard function to differ and be non-proportional across groups
   - Underlying baseline hazard is estimated for each stratum
   - Obtain one HR for each covariates
   - There should be no significant interactions between covariate and stratum variable

e.g. HR for smokers vs. non-smokers=1.7

Option B in Stata

Specify a ‘strata’ variable in st cox command

xi: stcox i.agegrp i.empgrade, strata(sex)

```
Stratified Cox regr. -- Breslow method for ties

No. of subjects =        10297                     Number of obs   =      10297
No. of failures =          605
Time at risk    =  176515.5127
                                                   LR chi2(5)      =     225.30
Log likelihood  =   -5024.4458                     Prob > chi2     =     0.0000

------------------------------------------------------------------------------
          _t | Haz. Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
  _Iagegrp_2 |   1.292502   .1976919     1.68   0.093      .957718    1.744315
  _Iagegrp_3 |   2.291222   .3289398     5.77   0.000     1.729276    3.035779
  _Iagegrp_4 |   4.105473   .5262094    11.02   0.000     3.193467    5.277934
_Iempgrade_2 |   1.306777   .1365392     2.56   0.010     1.064788    1.603761
_Iempgrade_3 |   1.966546   .2470575     5.38   0.000     1.537332    2.515595
------------------------------------------------------------------------------
                                                             Stratified by sex
```

When proportional hazards assumption is not met

- You may be able to split the follow up time
  - E.g. first 5 years, next 5-10 years – if hazards proportional within these bands

- It may also be possible to use a different survival regression method – a parametric survival model – but we will not cover this today

Cox regression assumptions

- Censoring should be independent of event of interest (non-informative censoring)

- All cause mortality:
  - Assumes end of study date not related to mortality

- Incident disease ascertained from follow-up visits
  - Assumes drop out from study is not related to incident CHD (this assumption might not hold as those that drop out might have worse health)

Censoring assumption

- If censoring is NOT independent of event of interest, then censoring is said to be informative

- You have to judge whether the assumption of non-informative censoring is met
  - as usually you will not have any data that allows you to test whether censoring is informative or non-informative

Choosing censoring dates

- Bear in mind that censoring should be independent of event

- Choosing start and end dates and coding survival times can be complicated
  - ➢ so always think carefully
  - ➢ AND always check a sample of records

Cox or logistic regression?

- May produce similar results for short or fixed follow-up periods (e.g. years for everyone)
- Results may differ if have varying follow up times
- If you have dates of entry and dates of events, it is better to use Cox regression

Time varying covariates

For example:

AGE: use current age group rather than age at baseline

EMPLOYMENT STATUS: may change over time

Time varying covariate - Example

- A person is employed for 2 years then becomes unemployed. They have a CHD event 6 months later. Employment status is a time-varying covariate

- To analyse this as a time varying covariate, the person’s record needs to be split into 2 records in the data – one for when they are employed and one for when they are unemployed

id	Emp status	time	event
10	1	2.0	0
10	2	0.5	1

Extensions to survival analysis

- Discrete (interval-censored) survival times
- Repeated events
- Multi-state models (more than 1 event type)
  - Incident disease and death
  - Coronary bypass and CHD event

Summary

- Time to event data can be analysed using Cox proportional hazards regression models

- These semi-parametric regression models provide estimates of the hazard ratio for covariates included in the model

- Cox regression models assume that hazards are proportional over the follow-up period

Useful website

- http://www.stats.ats.ucla.edu/stat/ has examples of survival analysis using STATA (and other packages)

Further reading

- Clayton D, Hills M. Statistical methods in epidemiology. Oxford University Press 1993

- Collet D. Modelling survival data in medical research. Texts in statistical science series. CC Press 2003

- Singer JD, Willett JB. Applied longitudinal data analysis. OUP 2003