Survival Analysis, Lecture 3

Parametric regression modelling

Manuela Quaresma

Department of Non-Communicable Disease Epidemiology
London School of Hygiene and Tropical Medicine

Aims

Part 1

- Explain why regression modelling for survival data is useful
- Formulate the likelihoods for survival data using parametric models, including the effects of binary explanatory variables
- Learn how to quantify the effect of explanatory variables

Part 2

- Define models using two parametric distributions for survival times: the exponential and Weibull distributions

Part 3

- Learn how to compare the fit of different models
- Extending beyond binary explanatory variables

Part 1: Introduction to parametric regression models for survival data

Session 2...Why use non-parametric methods

Non-parametric methods are a relatively simple starting point for most analyses of survival data.

- estimating survival functions and cumulative hazards
- provide a nice way of graphically displaying survival data
- making comparisons between two or more groups of individuals

Drawback of non-parametric methods

- they do not quantify the association between exposures and survival
- If we wish to adjust for potential confounders, we have to look separately at groups defined by the confounder and the methods quickly become cumbersome and the groups too small for meaningful analysis
- they do not allow us to investigate the impact of continuous variables on survival (e.g. blood pressure), unless we categorise and compare the hazards between different groups

Parametric regression modelling for survival data

- Regression modelling for survival data we assume a model for the survival times which includes how survival times depend on individual exposures (parametric).
- Estimate the effects of exposures
- Regression modelling for survival data is analogous to the use of linear regression (continuous responses) or logistic regression (binary outcomes)

Reminder of the likelihood for survival data

- For a censored individual i we observe that the person survived at least up until time t_i
- The probability that their unobserved survival time is beyond t_i is the survivor function

$$S(t_i) = \Pr(T > t_i)$$

- δ_i is an indicator (δ_i=1 indicates an event; δ_i=0 indicates a censoring time)

Full likelihood for survival data with censoring

$$L = \prod_{\text{survival times}} f(t_{E_i}) \prod_{\text{censoring times}} S(t_{C_i})$$

$$L = \prod_{i=1}^{n} f(t_i)^{\delta_i} S(t_i)^{1-\delta_i}$$

Quantifying the effect size

In the non-parametric setting:

- Compare survivor curves in two groups of individuals
- Test for a difference between groups using the log rank test

However...

This does NOT quantify the effect of the explanatory variable on survival.

Quantifying the effect size

Considering a single binary explanatory variable X observed on each individual at the start of follow-up

$$X = \begin{cases} 0 \\ 1 \end{cases} \tag{1}$$

- In a randomized trial setting X it may refer to treatment group.
- In an observational study of an occupational cohort, X may refer to occupational exposure to radiation.
- In a population-based cohort X may refer to smoking status [smoker or non-smoker].

Quantifying the effect size

Assume the hazard in one group of individuals (X=1) is a multiple of the hazard in the 'baseline' group (X=0)

- $h_0(t)$, hazard function in the X=0 group
- $h_1(t)$, hazard function in the X=1 group

we can write

$$h_1(t) = \psi h_0(t)$$

where ψ is a parameter to be estimated.

Since the hazard cannot be negative, ψ cannot be negative.

Quantifying the effect size

For this reason it is convenient instead to write:

$$h_1(t) = e^{\beta} h_0(t), \text{ i.e. } \psi = e^{\beta}$$

using this formulation the parameter β can take any value

- This model is called proportional hazards model
- The assumption that β does not depend on t is called the proportional hazards assumption

The ratio of the hazards in the two groups is

$$\frac{h_1(t)}{h_0(t)} = e^{\beta} \tag{2}$$

e^{β} is called the hazard ratio, and β the log hazard ratio

$\Longrightarrow$ this ratio does not depend on time (t)

Quantifying the effect size

For this reason it is convenient instead to write:

$$h_1(t) = e^{\beta} h_0(t), \text{ i.e. } \psi = e^{\beta}$$

using this formulation the parameter β can take any value

- This model is called proportional hazards model
- The assumption that β does not depend on t is called the proportional hazards assumption

The ratio of the hazards in the two groups is

$$\frac{h_1(t)}{h_0(t)} = e^{\beta} \tag{2}$$

e^{β} is called the hazard ratio, and β the log hazard ratio

$\Longrightarrow$ this ratio does not depend on time (t)

Quantifying the effect size

For this reason it is convenient instead to write:

$$h_1(t) = e^{\beta} h_0(t), \text{ i.e. } \psi = e^{\beta}$$

using this formulation the parameter β can take any value

- This model is called proportional hazards model
- The assumption that β does not depend on t is called the proportional hazards assumption

The ratio of the hazards in the two groups is

$$\frac{h_1(t)}{h_0(t)} = e^{\beta} \tag{2}$$

e^{β} is called the hazard ratio, and β the log hazard ratio

$\Longrightarrow$ this ratio does not depend on time (t)

Part 2: The Exponential and the Weibull models

The exponential model

Hazard function is constant over time. Rate at which events occur is constant over the time scale

$$h(t) = \lambda, \qquad S(t) = e^{-\lambda t}, \qquad f(t) = \lambda e^{-\lambda t}$$

Note that as defined in Part 1: $h_1(t) = e^{\beta} h_0(t)$

To incorporate a binary exploratory variable X we write

$$\begin{cases} h(t;0) = \lambda, & X = 0 \\ h(t;1) = \lambda e^{\beta}, & X = 1 \end{cases} \tag{3}$$

More conveniently

$$h(t;x) = \lambda e^{\beta x}$$

The exponential model

Hazard function is constant over time. Rate at which events occur is constant over the time scale

$$h(t) = \lambda, \qquad S(t) = e^{-\lambda t}, \qquad f(t) = \lambda e^{-\lambda t}$$

Note that as defined in Part 1: $h_1(t) = e^{\beta} h_0(t)$

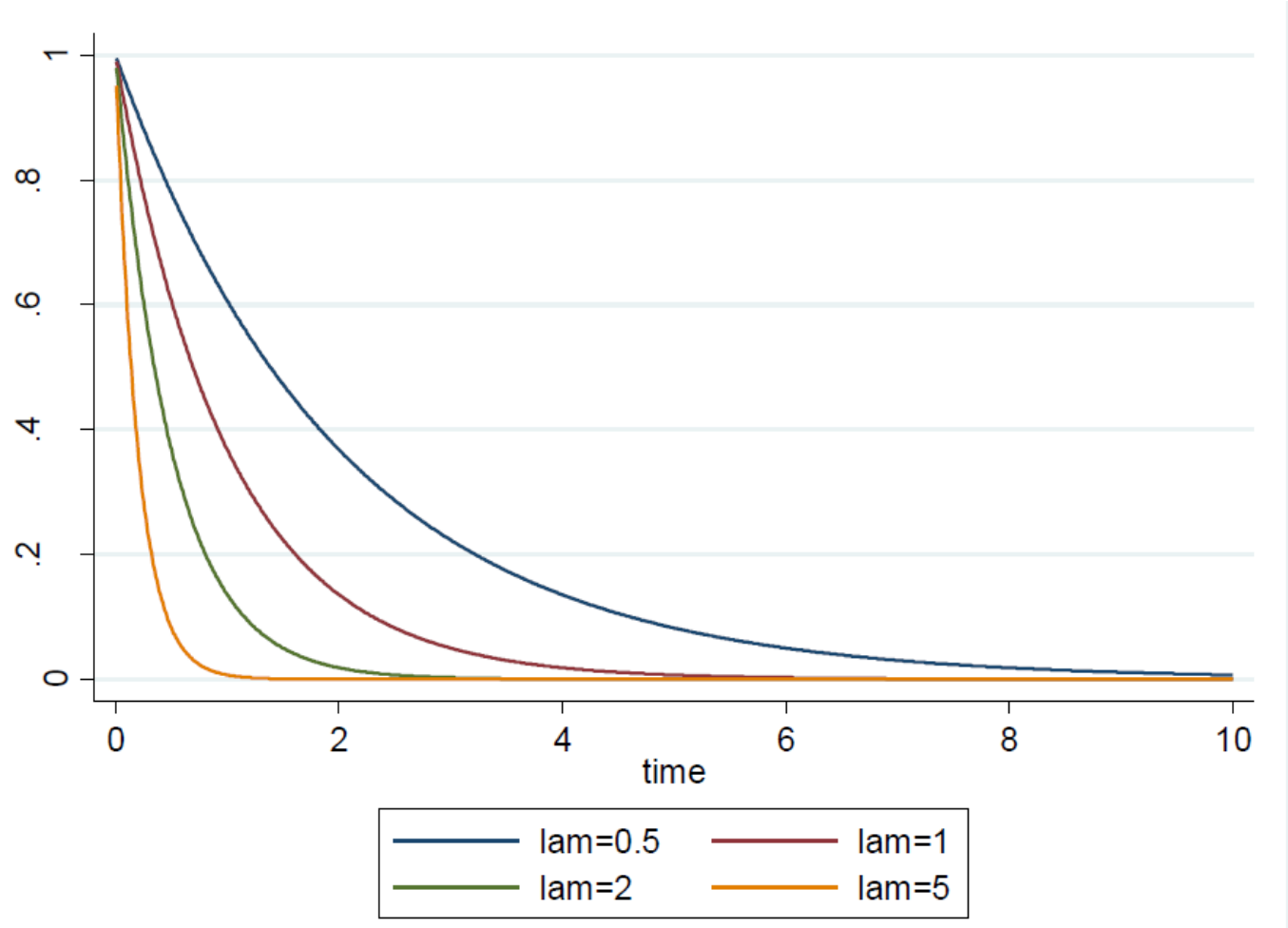
To incorporate a binary exploratory variable X we write

$$\begin{cases} h(t;0) = \lambda, & X = 0 \\ h(t;1) = \lambda e^{\beta}, & X = 1 \end{cases} \tag{3}$$

More conveniently

$$h(t;x) = \lambda e^{\beta x}$$

The exponential model - Survival functions



The exponential model

Probability density function and survivor function are

$f(t;x) = \lambda e^{\beta x} exp(-\lambda t e^{\beta x}), \qquad S(t;x) = exp(-\lambda t e^{\beta x})$

And the likelihood of the data

$L = \prod_{i=1}^{n} f(t_i)^{\delta_i} S(t_i)^{1-\delta_i}$

$L = \prod_{i=1}^{n} [\lambda e^{\beta x_i} exp(-\lambda t_i e^{\beta x_i})]^{\delta_i} [exp(-\lambda t_i e^{\beta x_i})]^{1-\delta_i}$

t_i survival or censoring time; δ_i event indicator; x_i exposure

MLEs for λ and β found in the usual way by differentiating the log likelihood with respect to both parameters

The exponential model

Probability density function and survivor function are

$f(t;x) = \lambda e^{\beta x} exp(-\lambda t e^{\beta x}), \qquad S(t;x) = exp(-\lambda t e^{\beta x})$

And the likelihood of the data

$L = \prod_{i=1}^{n} f(t_i)^{\delta_i} S(t_i)^{1-\delta_i}$

$L = \prod_{i=1}^{n} [\lambda e^{\beta x_i} exp(-\lambda t_i e^{\beta x_i})]^{\delta_i} [exp(-\lambda t_i e^{\beta x_i})]^{1-\delta_i}$

t_i survival or censoring time; δ_i event indicator; x_i exposure

MLEs for λ and β found in the usual way by differentiating the log likelihood with respect to both parameters

The exponential model

Perform a test of the null hypothesis that the hazard ratio is 1, i.e. there is no difference between the hazard rates in the two groups

Assuming a 2-sided alternative hypothesis:

$$\begin{aligned} \text{Null hypothesis:} \quad & e^{\beta} = 1 \\ \text{Alternative hypothesis:} \quad & e^{\beta} \neq 1 \end{aligned}$$

or equivalently

$$\begin{aligned} \text{Null hypothesis:} \quad & \beta = 0 \\ \text{Alternative hypothesis:} \quad & \beta \neq 0 \end{aligned}$$

A test of the null hypothesis is obtained using the Wald test

$$\frac{\widehat{\beta}}{SE(\widehat{\beta})} \sim N(0,1)$$

The exponential model: (example 3.1)

Times to death in leukaemia patients in two groups: control (X=0) and treatment (X=1)

Table 3.1. Results from fitting an exponential model to the leukaemia data.

Parameter	Estimate	Standard error	95% confidence interval	p-value
λ	0.12	0.03	(0.08,0.18)	<0.001
β	-1.53	0.40	(-2.31,-0.75)	<0.001
$\exp \beta$	0.22	0.09	(0.10,0.48)	<0.001

- Estimated hazard ratio=0.22 (95%CI 0.10, 0.48)
- Hazard rate for death in treatment group is 0.22 (78% reduction) that of the control group
- λ gives the hazard in the control group

The exponential model: (example 3.1)

Times to death in leukaemia patients in two groups: control (X=0) and treatment (X=1)

Table 3.1. Results from fitting an exponential model to the leukaemia data.

Parameter	Estimate	Standard error	95% confidence interval	p-value
λ	0.12	0.03	(0.08,0.18)	<0.001
β	-1.53	0.40	(-2.31,-0.75)	<0.001
$\exp \beta$	0.22	0.09	(0.10,0.48)	<0.001

- Estimated hazard ratio=0.22 (95%CI 0.10, 0.48)
- Hazard rate for death in treatment group is 0.22 (78% reduction) that of the control group
- λ gives the hazard in the control group

The Weibull distribution

In many applications it will not be reasonable to assume a constant hazard rate over time $\Longrightarrow$ alternative Weibull distribution

Hazard function

$h(t) = k\lambda t^{k-1}$ (compare to exponential: $h(t) = \lambda$)

Survival function

$S(t) = exp(-\lambda t^k)$ (compare to exponential: $S(t) = exp(-\lambda t)$)

k=shape parameter; λ=scale parameter

- Values of $k > 1$ indicate a hazard increasing over time
- Values of $k < 1$ indicate a hazard decreasing over time

The Weibull distribution

To incorporate a binary explanatory variable X, assuming a proportional hazard situation

$$h(t;x) = k\lambda t^{k-1} e^{\beta x}$$

$$S(t;x) = exp(-\lambda t^{k} e^{\beta x})$$

The likelihood of the data

$$L = \prod_{i=1}^{n} f(t_i)^{\delta_i} S(t_i)^{1-\delta_i}$$

$$L = \prod_{i=1}^{n} [k\lambda t_i^{k-1} e^{\beta x_i} exp(-\lambda t_i^{k} e^{\beta x_i})]^{\delta_i} [exp(-\lambda t_i^{k} e^{\beta x_i})]^{1-\delta_i}$$

The Weibull distribution

To incorporate a binary explanatory variable X, assuming a proportional hazard situation

$$h(t;x) = k\lambda t^{k-1} e^{\beta x}$$

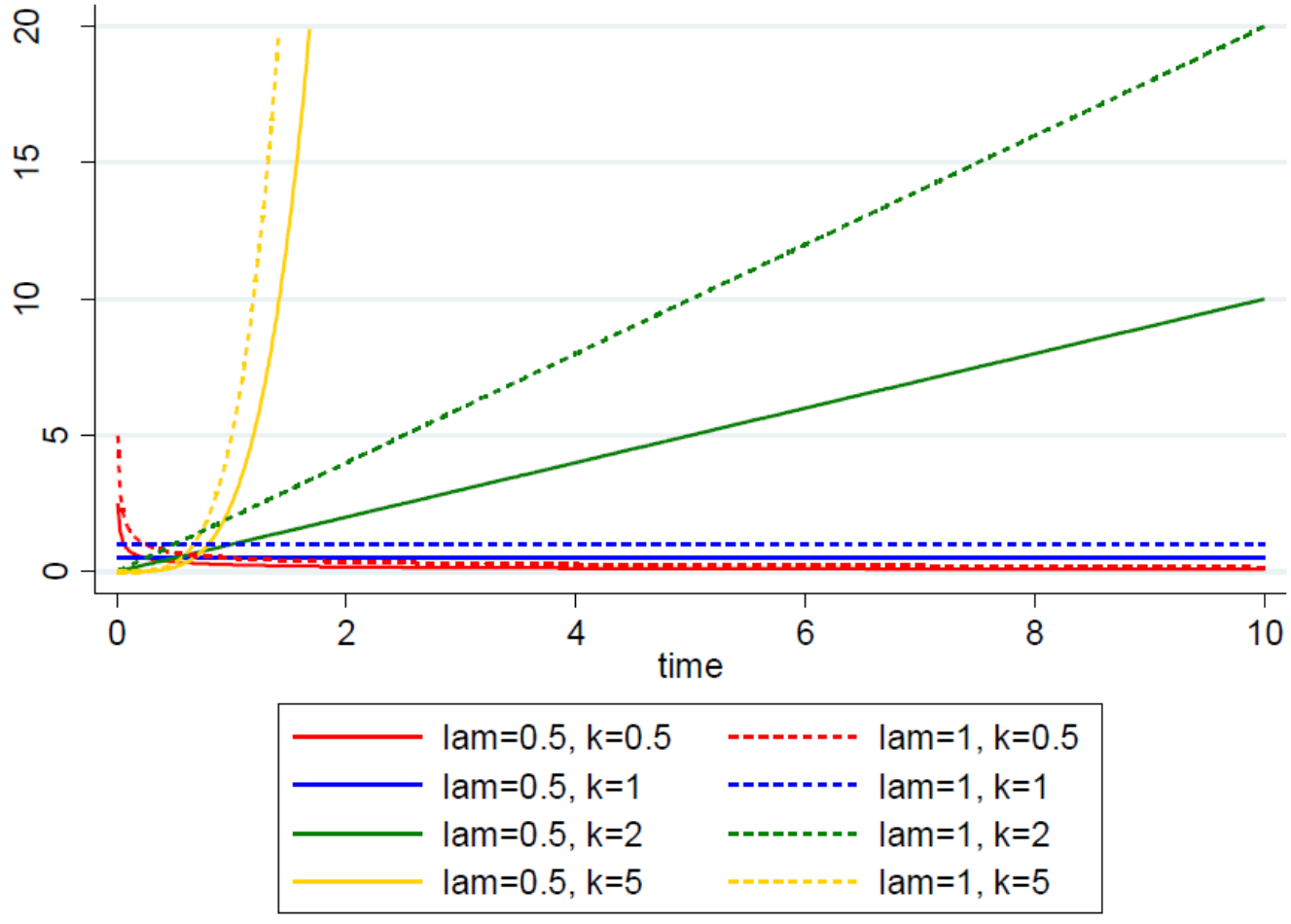
$$S(t;x) = exp(-\lambda t^{k} e^{\beta x})$$

The likelihood of the data

$$L = \prod_{i=1}^{n} f(t_i)^{\delta_i} S(t_i)^{1-\delta_i}$$

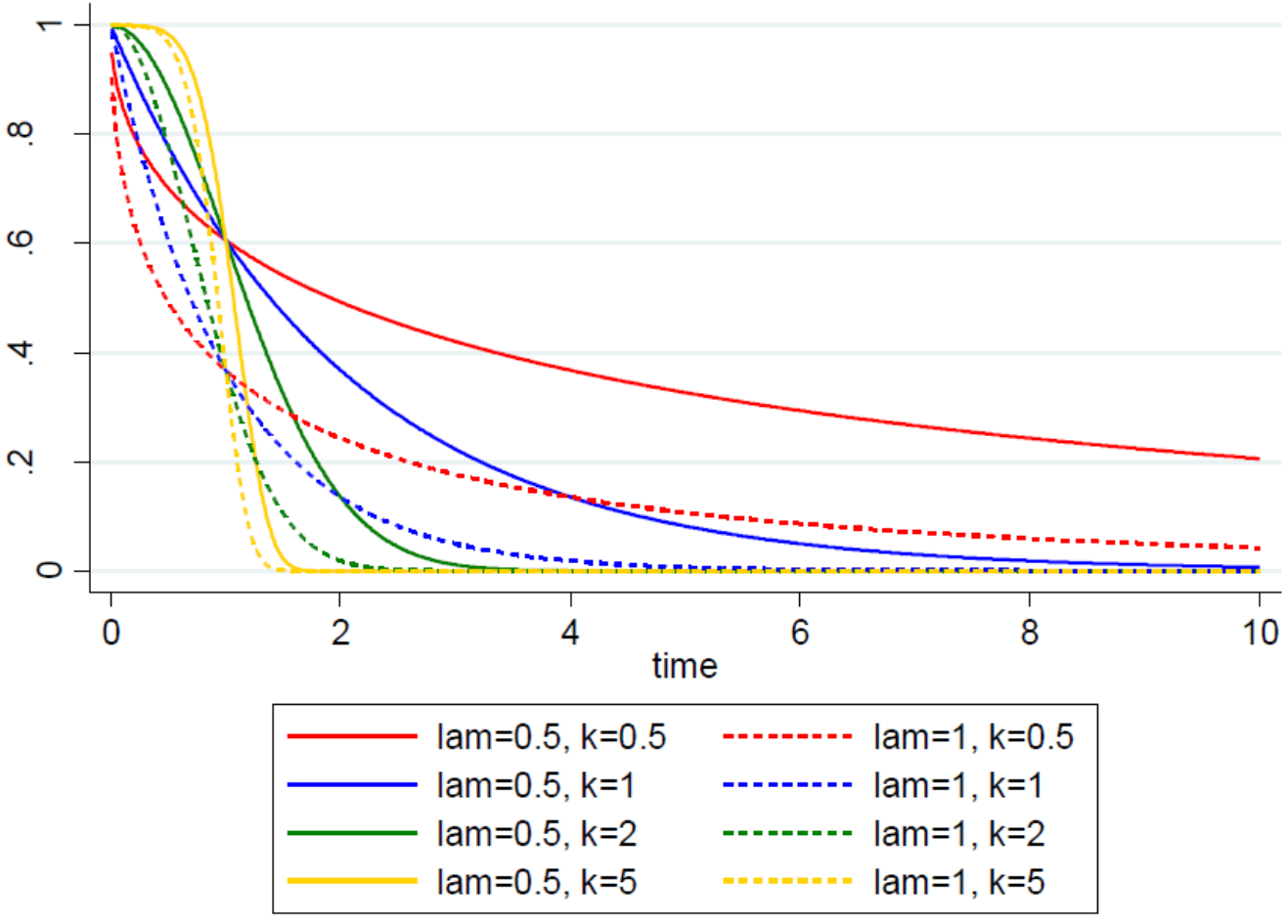
$$L = \prod_{i=1}^{n} [k\lambda t_i^{k-1} e^{\beta x_i} exp(-\lambda t_i^{k} e^{\beta x_i})]^{\delta_i} [exp(-\lambda t_i^{k} e^{\beta x_i})]^{1-\delta_i}$$

The Weibull model - Hazard functions



Monotonically increasing ($k > 1$) or decreasing ($k < 1$)

The Weibull model - Survival functions



The Weibull model: (example 3.2)

Times to death in leukaemia patients in two groups: control (X=0) and treatment (X=1)

Table 3.2. Results from fitting a Weibull model to the leukaemia patient data.

Parameter	Estimate	Standard error	95% confidence interval	p-value
λ	0.05	0.03	(0.02,0.14)	<0.001
κ	1.37	-	(1.02,1.82)	0.034*
β	-1.73	0.41	(-2.54,-0.92)	<0.001
$\exp \beta$	0.18	0.07	(0.08,0.40)	<0.001

- Estimated hazard ratio=0.18 (95%CI 0.08, 0.40)
- Similar to the hazard ratio=0.22 from the exponential model
- * This is the p-value for a test of logk = 0

The Weibull model: (example 3.2)

Times to death in leukaemia patients in two groups: control (X=0) and treatment (X=1)

Table 3.2. Results from fitting a Weibull model to the leukaemia patient data.

Parameter	Estimate	Standard error	95% confidence interval	p-value
λ	0.05	0.03	(0.02,0.14)	<0.001
κ	1.37	-	(1.02,1.82)	0.034*
β	-1.73	0.41	(-2.54,-0.92)	<0.001
$\exp \beta$	0.18	0.07	(0.08,0.40)	<0.001

- Estimated hazard ratio=0.18 (95%CI 0.08, 0.40)
- Similar to the hazard ratio=0.22 from the exponential model
- * This is the p-value for a test of logk = 0

Part 3:

- Comparing the fit of Exponential and Weibull models
- Extending beyond binary explanatory variables

Comparing the fit of the exponential and Weibull models

Important question: how can we choose a good parametric model for our data?

Two ways of assessing whether exponential or Weibull model is appropriate:

- Using plots
- Using statistical tests

Comparing the fit of Exponential and Weibull models

Using plots

Plotting non-parametric estimates of survival can give an indication as to whether exponential or Weibull models are suitable

- Under an exponential distribution the cumulative hazard is

$$H(t;x) = -logS(t;x) = \lambda t e^{\beta x}$$

$\Longrightarrow$ cumulative hazard is linear in t in both exposure groups

- Under a Weibull distribution

$$logH(t;x) = log(-logS(t;x)) = log\lambda + k logt + \beta x$$

$\Longrightarrow$ log cumulative hazard is linear in $logt$ with constant shift between the groups

Comparing the fit of Exponential and Weibull models

Using plots

Plotting non-parametric estimates of survival can give an indication as to whether exponential or Weibull models are suitable

- Under an exponential distribution the cumulative hazard is

$$H(t;x) = -logS(t;x) = \lambda t e^{\beta x}$$

$\Longrightarrow$ cumulative hazard is linear in t in both exposure groups

- Under a Weibull distribution

$$logH(t;x) = log(-logS(t;x)) = log\lambda + k logt + \beta x$$

$\Longrightarrow$ log cumulative hazard is linear in $logt$ with constant shift between the groups

Comparing the fit of Exponential and Weibull models

Using plots

Plotting non-parametric estimates of survival can give an indication as to whether exponential or Weibull models are suitable

- Under an exponential distribution the cumulative hazard is

$$H(t;x) = -logS(t;x) = \lambda t e^{\beta x}$$

$\Longrightarrow$ cumulative hazard is linear in t in both exposure groups

- Under a Weibull distribution

$$logH(t;x) = log(-logS(t;x)) = log\lambda + k\,logt + \beta x$$

$\Longrightarrow$ log cumulative hazard is linear in $logt$ with constant shift between the groups

Comparing the fit of Exponential and Weibull models

- Under an exponential distribution

  A plot of $H(t;x)$ against t should be linear within groups

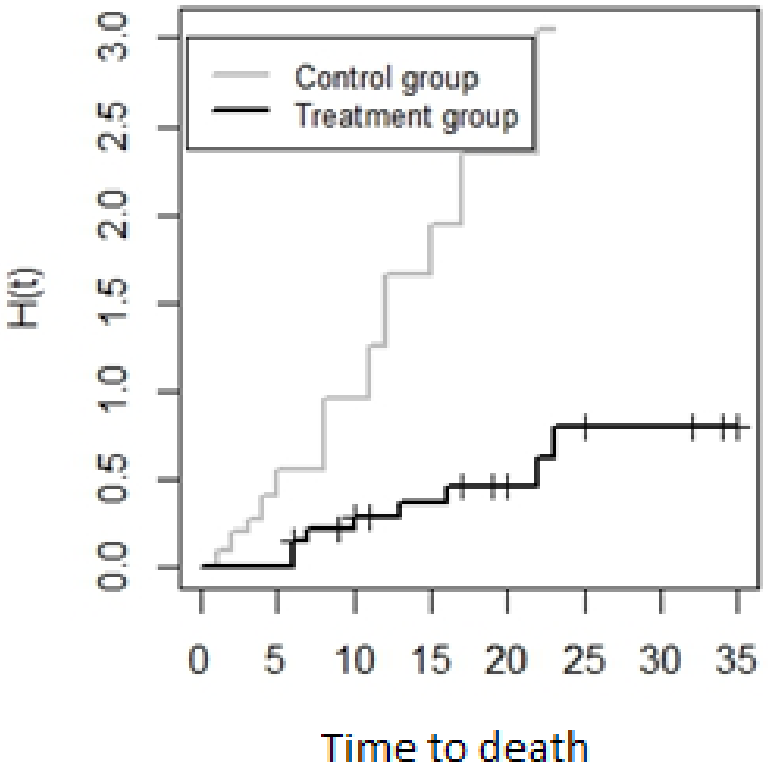
- Under a Weibull distribution

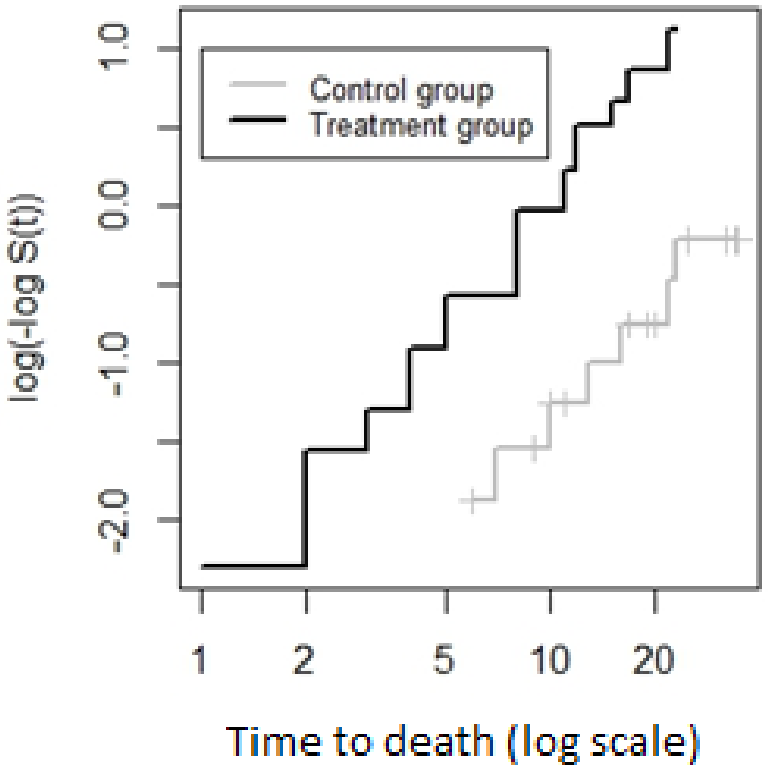
  A plot of $log(H(t;x))$ against log t should be linear within groups

This can be investigated using Kaplan-Meier plots

Investigating the exponential and Weibull models for the leukaemia patient data

Exponential or Weibull?





Comparing the fit of Weibull and exponential models

Using statistical tests

Exponential model special case of Weibull model with k=1

A test of the null hypothesis that the hazard is constant over time is a test of H0: k=1 (or logk=0)

- Weibull model: $h(t;x) = k\lambda t^{k-1} e^{\beta x}$

  reduces to (when k=1)

- Exponential model: $h(t;x) = \lambda e^{\beta x}$

Comparing the fit of Weibull and exponential models

Testing the hypothesis H0: k=1 (or logk=0)

1. Wald test:
   - performed using the estimate of log k and its standard error
   - The test statistics is compared with the standard Normal distribution

$$\frac{\widehat{logk}}{SE(\widehat{logk})} \sim N(0,1) \tag{4}$$

2. Likelihood ratio test (LRT):
   - compare likelihoods from Exponential model and Weibull model
   - Exponential model is nested within Weibull model

$$-2(\ell_{exponential} - \ell_{Weibull}) \sim \chi^2_1 \tag{5}$$

Degrees of freedom: difference in the number of estimated parameters
LRT more powerful than the Wald test and is preferred

Comparing the fit of Weibull and exponential models

Testing the hypothesis H0: k=1 (or logk=0)

1. Wald test:
   - performed using the estimate of log k and its standard error
   - The test statistics is compared with the standard Normal distribution

$$\frac{\widehat{logk}}{SE(\widehat{logk})} \sim N(0,1) \tag{4}$$

2. Likelihood ratio test (LRT):
   - compare likelihoods from Exponential model and Weibull model
   - Exponential model is nested within Weibull model

$$-2(\ell_{exponential} - \ell_{Weibull}) \sim \chi^2_1 \tag{5}$$

Degrees of freedom: difference in the number of estimated parameters
LRT more powerful than the Wald test and is preferred

Comparing the fit of Weibull and exponential models

Testing the hypothesis H0: k=1 (or logk=0)

1. Wald test:
   - performed using the estimate of log k and its standard error
   - The test statistics is compared with the standard Normal distribution

$$\frac{\widehat{logk}}{SE(\widehat{logk})} \sim N(0,1) \tag{4}$$

2. Likelihood ratio test (LRT):
   - compare likelihoods from Exponential model and Weibull model
   - Exponential model is nested within Weibull model

$$-2(\ell_{exponential} - \ell_{Weibull}) \sim \chi^2_1 \tag{5}$$

Degrees of freedom: difference in the number of estimated parameters
LRT more powerful than the Wald test and is preferred

Investigating the exponential and Weibull models for the leukaemia patient data

What is the interpretation of p=1.37?

```
. streg group, distribution(exponential)
                                                LR chi2(1)      =      16.49
Log likelihood  =    -49.00866                  Prob > chi2     =     0.0000

------------------------------------------------------------------------------
          _t | Haz. Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       group |   .2172702   .0865625    -3.83   0.000     .0995115    .4743806
       _cons |   .1153846    .025179    -9.90   0.000     .0752316    .1769682
------------------------------------------------------------------------------
. streg group, distribution(weibull)
                                                LR chi2(1)      =      19.65
Log likelihood  =   -47.064102                  Prob > chi2     =     0.0000

------------------------------------------------------------------------------
          _t | Haz. Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       group |   .1771299   .0731691    -4.19   0.000     .0788272    .3980227
       _cons |   .0463885    .025888    -5.50   0.000     .0155375     .138497
-------------+----------------------------------------------------------------
      /ln_p  |   .3117092   .1472919     2.12   0.034     .0230224     .600396
-------------+----------------------------------------------------------------
           p |   1.365757    .201165                      1.02329     1.82284
         1/p |   .7321944   .1078463                     .5485944    .9772406
```

Note: The ”k” parameter is called ”p” in STATA output

Investigating the exponential and Weibull models for the leukaemia patient data

What is the interpretation of lnp=0.31 and p-value=0.034?

```
. streg group, distribution(exponential)
                                                LR chi2(1)      =      16.49
Log likelihood  =    -49.00866                  Prob > chi2     =     0.0000

------------------------------------------------------------------------------
          _t | Haz. Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       group |   .2172702   .0865625    -3.83   0.000     .0995115    .4743806
       _cons |   .1153846    .025179    -9.90   0.000     .0752316    .1769682
------------------------------------------------------------------------------
. streg group, distribution(weibull)
                                                LR chi2(1)      =      19.65
Log likelihood  =   -47.064102                  Prob > chi2     =     0.0000

------------------------------------------------------------------------------
          _t | Haz. Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       group |   .1771299   .0731691    -4.19   0.000     .0788272    .3980227
       _cons |   .0463885    .025888    -5.50   0.000     .0155375     .138497
-------------+----------------------------------------------------------------
      /ln_p  |   .3117092   .1472919     2.12   0.034     .0230224     .600396
-------------+----------------------------------------------------------------
           p |   1.365757    .201165                      1.02329     1.82284
         1/p |   .7321944   .1078463                     .5485944    .9772406
```

Note: The ”k” parameter is called ”p” in STATA output

Investigating the exponential and Weibull models for the leukaemia patient data

```
> leukaemia.weib<- weibreg(Surv(time=time,event=death)~as.factor(group), data=leukaemia)
> leukaemia.weib
Call:
weibreg(formula = Surv(time = time, event = death) ~ as.factor(group),
    data = leukaemia)

Covariate           Mean       Coef Exp(Coef)  se(Coef)    Wald p
as.factor(group)
               0    0.336     0          1           (reference)
               1    0.664    -1.731      0.177     0.413     0.000

log(scale)                    2.248      9.472     0.166     0.000
log(shape)                    0.312      1.366     0.147     0.034

Events
Total time at risk           541
Max. log. likelihood      -106.58
LR test statistic         19.6
Degrees of freedom        1
Overall p-value           9.29141e-06
```

Note: The ”k” parameter is called ”shape” in R output

Extending beyond binary explanatory variables

Methods outlined for a binary variable can be extended to:

- continuous explanatory variables
- categorical explanatory variables
- multiple explanatory variables

For a vector of explanatory variables $X = (X_1, X_2, \ldots, X_p)^T$ the proportional hazards assumption is

$$h(t; x) = h_0(t) e^{\beta^T x}$$

where $h_0(t)$ is the baseline hazard and $\beta = (\beta_1, \beta_2, \ldots, \beta_p)^T$ is a vector of parameters to be estimated

We outline the interpretation of β in different circumstances

Continuous explanatory variable X

The effect of an increase of 1 unit in the continuous variable X is to multiply the hazard by e^{β}

Ratio of the hazards for a person with X=1 and a person with X=0 is

$$\frac{h(t;1)}{h(t;0)} = \frac{h_0(t)e^{\beta}}{h_0(t)} = e^{\beta} \tag{6}$$

For a continuous variable X, β is the log hazard ratio associated with a 1 unit increase in X.

Categorical explanatory variable X with more than 2 categories

For a categorical variable with K+1 categories, we define a series of indicator variables

$$X_k = \begin{cases} 1 & \text{if in category k} \\ 0 & \text{otherwise} \end{cases} \tag{7}$$

The hazard in category k is

$$h(t|X_k = 1) = h_0(t)e^{\beta_k}, k = 0, 1, \ldots, K$$

ou equivalently,

$$h(t|x_1, ..., x_K) = h_0(t)exp(\beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_K x_K)$$

- The baseline hazard refers to an individual in the baseline category (assumed to be $X_0 = 1$). It is assumed that $\beta_0 = 0$.
- e^{β_k} is the hazard ratio which compares individuals in category k with individuals in category 0

More than one explanatory variable

In general if we have a vector of explanatory variables $X = (X_1, X_2, \ldots, X_p)^T$:

- binary
- categorical
- continuous

and $\beta = (\beta_1, \beta_2, \ldots, \beta_p)^T$ a vector of corresponding log hazard ratio estimates.

- The interpretation of a particular element of β is as the log hazard ratio for a particular variable, holding all other elements of X fixed.
- That is, β_k, say, is the log hazard ratio for a unit increase in X_k conditional on all of the other variables in X.