Lecture 16
Regression with Time-to-event outcomes

BIOST 515

March 2, 2004

Outline

- Parametric models
  - Proportional hazards
  - Accelerated failure time
- Cox proportional hazards

Regression

In linear regression, we related a set of predictors to the outcome through

$$E(Y_i) = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip}.$$

How do we interpret β_k, $= 1, \ldots, p$?

In logistic regression, we related a set of predictors or risk factors to the outcome through

$$\text{logit}(\pi_i) = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip}.$$

How do we interpret β_k, $= 1, \ldots, p$?

Survival regression

There are several ways to relate the outcome to predictors in survival analysis. We will focus on two

- Proportional hazards (relative risk)

$$h(t|x) = h(t)\exp(\beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip})$$

- Accelerated failure time

$$S(t|X) = \psi((\log(t) - X\beta)/\sigma),$$

where ψ is any standardized survival distribution.

Proportional hazards (relative risk)

- The most widely used survival regression specification.
- Predictors act on a subject's hazard.
- The form of the regression is

$$h(t|X) = h(t)\exp(X\beta),$$

  where $h(t)$ is referred to as an $underlying\ hazard\ function$.

- Any parametric hazard function can be used for $h(t)$.
  - Later, we will see that $h(t)$ can be left completely unspecified.

- Depending on parametric form, $X\beta$ may have an intercept.
- The term $\exp(X\beta)$ is sometimes called a *relative hazard function*.
- The PH model can be linearized with respect to $X\beta$ using the following identities

$$\log h(t|X) = \log h(t) + X\beta$$

$$\log H(t|X) = \log H(t) + X\beta$$

Assumptions for a parametric PH model

- The true form of the underlying functions ($h,\ H,\ S$) are specified correctly.
- The relationship between the predictors and the log hazard is linear.
- In the absence of interactions, the predictors act additively on the log hazard.
- The effect of the predictors is the same for all values of t.

Interpretation of coefficients

The regression coefficient for X_j is the increase in log hzard at any fixed point in time if X_j is increased by one unit and all other predictors are held constant.

$$\begin{aligned}\beta_j &= \log h(t|X_1, X_2, \ldots, X_j + 1, X_{j+1}, \ldots, X_p) - \\ &\qquad \log h(t|X_1, X_2, \ldots, X_j, X_{j+1}, \ldots, X_p) \\ &= \log \frac{h(t|X_1, X_2, \ldots, X_j + 1, X_{j+1}, \ldots, X_p)}{h(t|X_1, X_2, \ldots, X_j, X_{j+1}, \ldots, X_p)}.\end{aligned}$$

This translates to

$$\exp(\beta_j) = \frac{h(t|X_1, X_2, \ldots, X_j + 1, X_{j+1}, \ldots, X_p)}{h(t|X_1, X_2, \ldots, X_j, X_{j+1}, \ldots, X_p)}.$$

How do we interpret $\exp(\beta_j)$?

The effect of increasing X_j by 1 is to increase the hazard of the event by a factor of $\exp(\beta_j)$ at all points in time.

What if we increase X_j by Δ?

In general the ratio of hazards for an individual with predictor values X^* compared to an individual with predictor values X is

$$\begin{aligned} \text{hazard ratio}(X^* : X) &= \frac{h(t)\exp(X^*\beta)}{h(t)\exp(X\beta)} \\ &= \frac{\exp(X^*\beta)}{\exp(X\beta)} = \exp[(X^* - X)\beta]. \end{aligned}$$

Example with one binary predictor

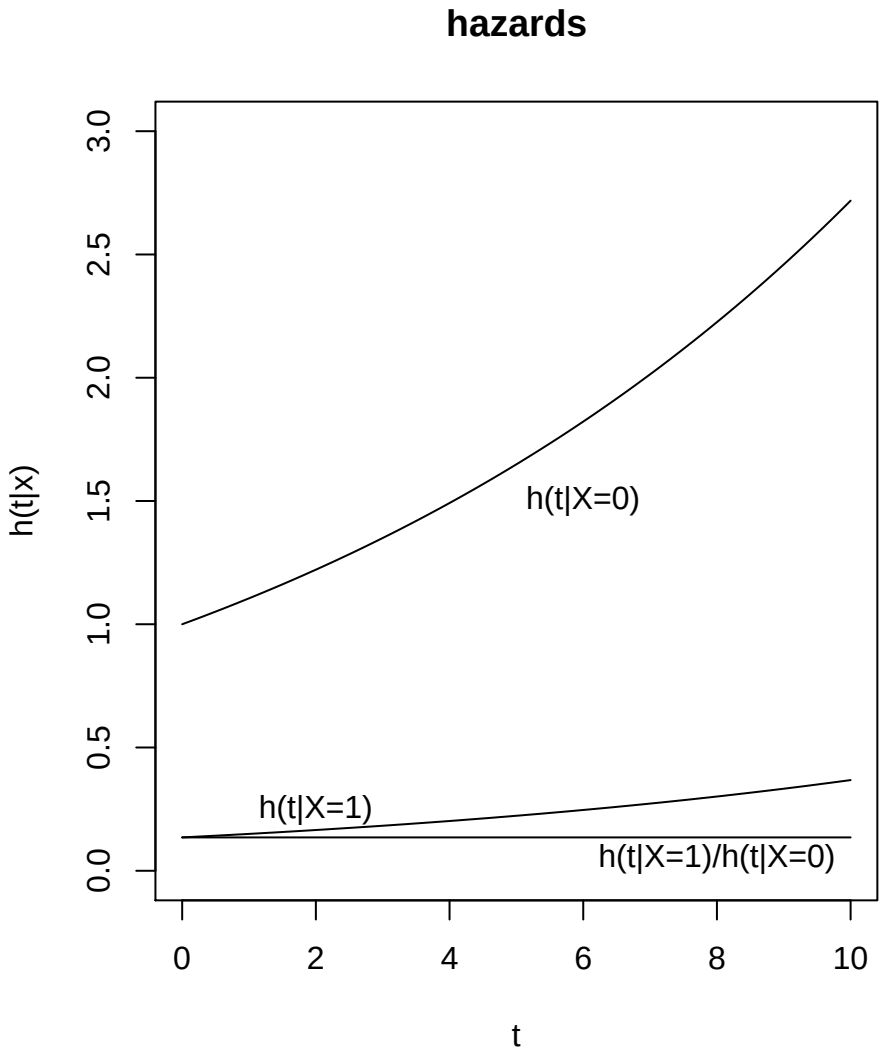
- X_1 is a binary predictor
  - sex: $X_1 = 1$ if subject is male, $X_1 = 0$ if subject is female.
  - treatment: $X_1 = 1$ if subject is on active treatment, $X_1 = 0$ if subject is on placebo
  - risk factor: $X_1 = 1$ if risk factor is present, $X_1 = 0$ if not

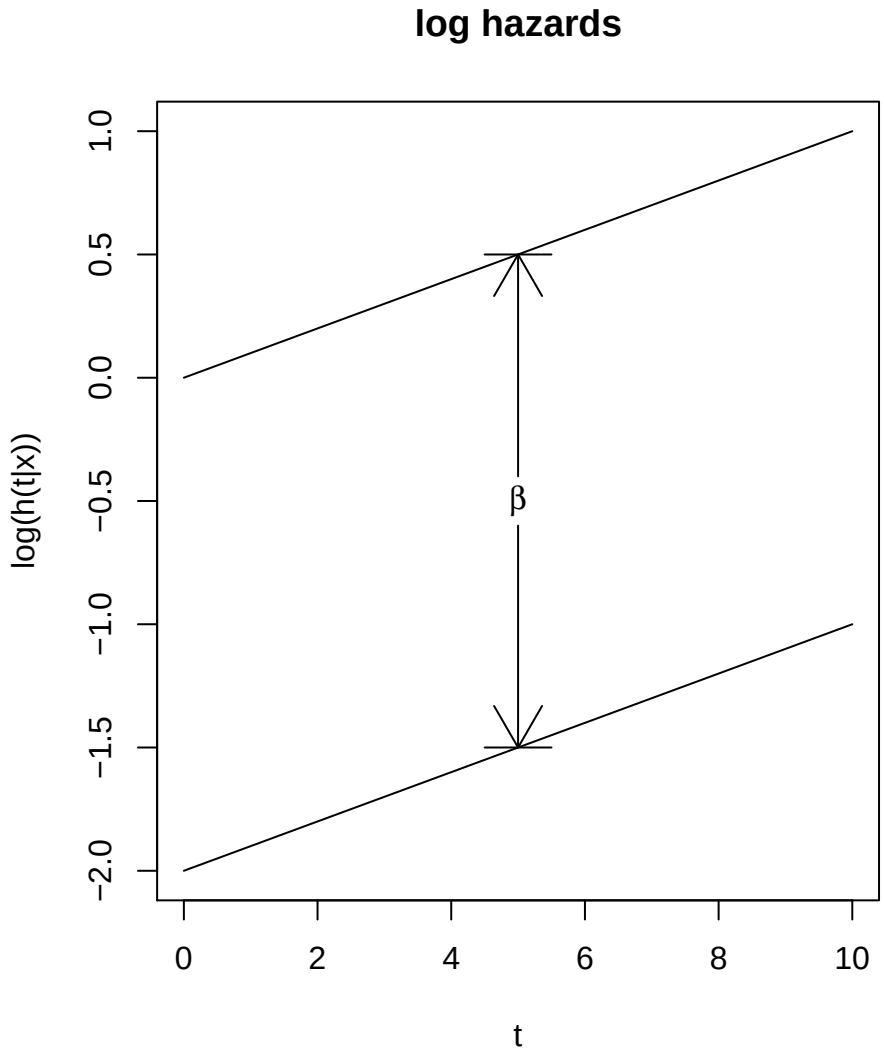
- The PH model (without intercept) can be written

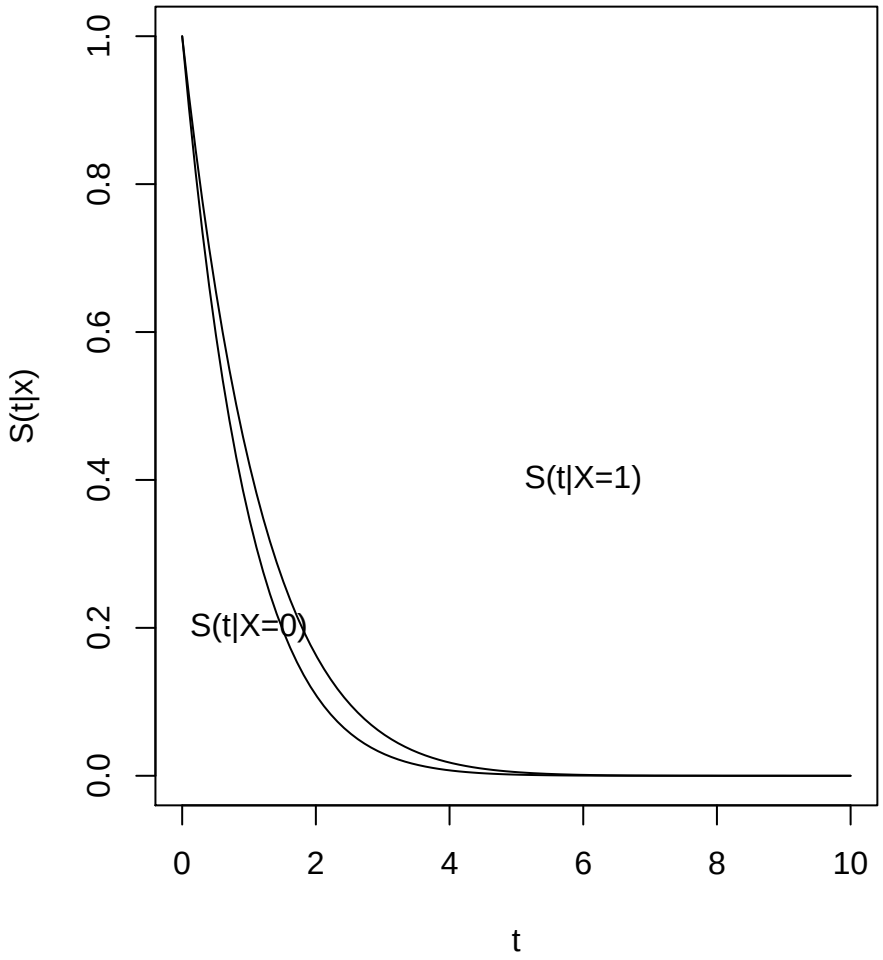
$$h(t|X_1 = 0) = h(t)$$

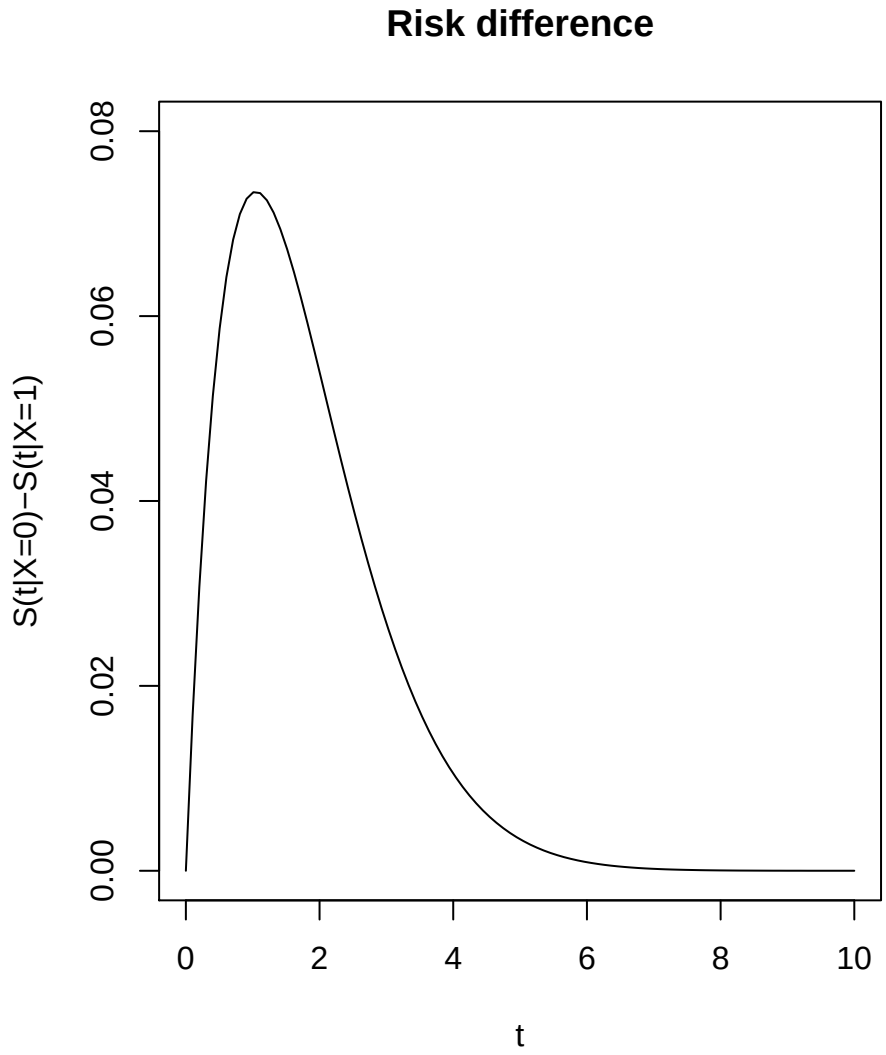
$$h(t|X_1 = 1) = h(t)\exp(\beta_1).$$

- hr$(X_1 = 1 : X_1 = 0) = \exp(\beta_1)$.

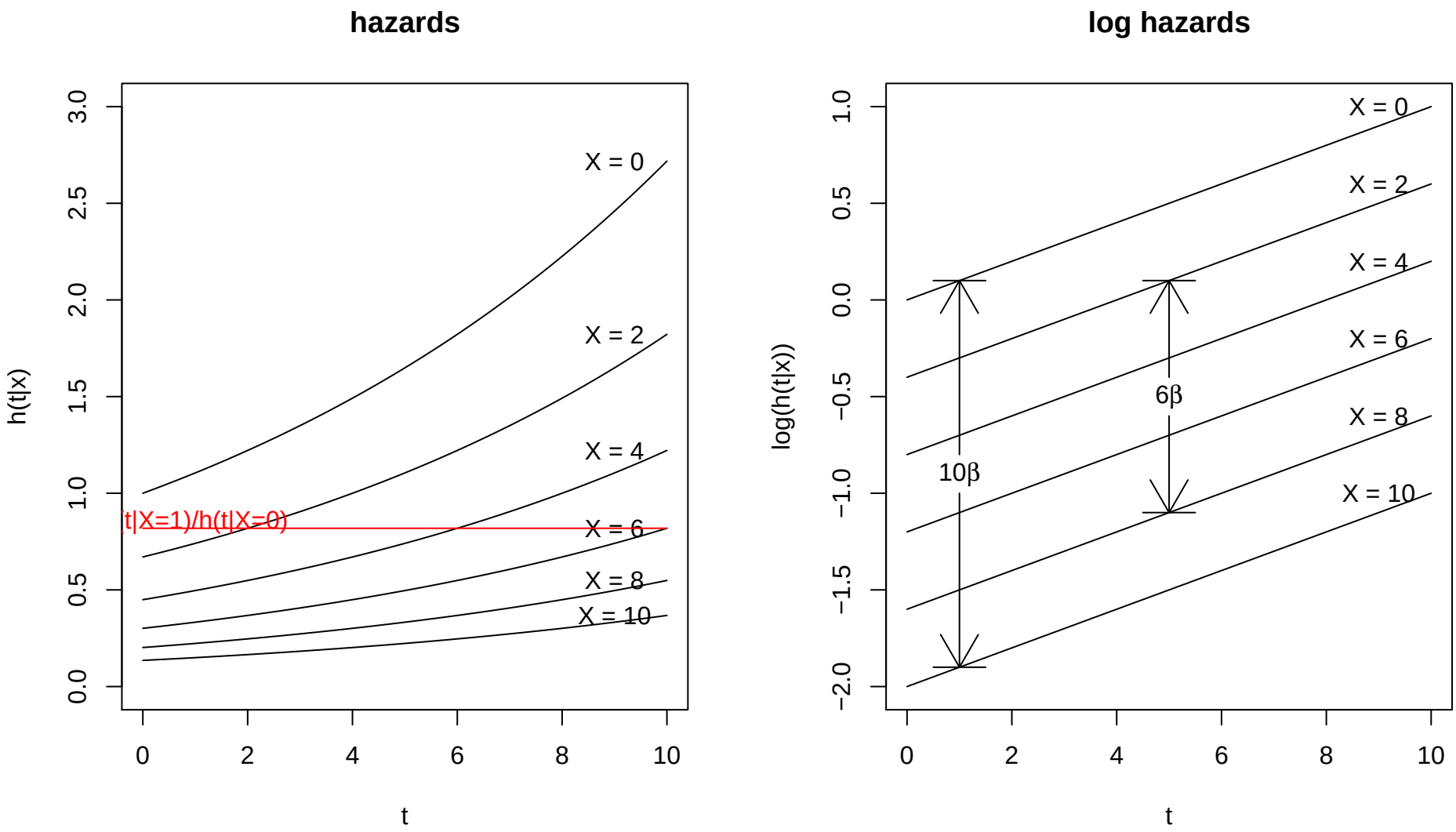

hazards
h(t|x)
3.0
2.5
2.0
1.5
1.0
0.5
0.0
h(t|X=0)
h(t|X=1)
h(t|X=1)/h(t|X=0)
0
2
4
6
8
10
t

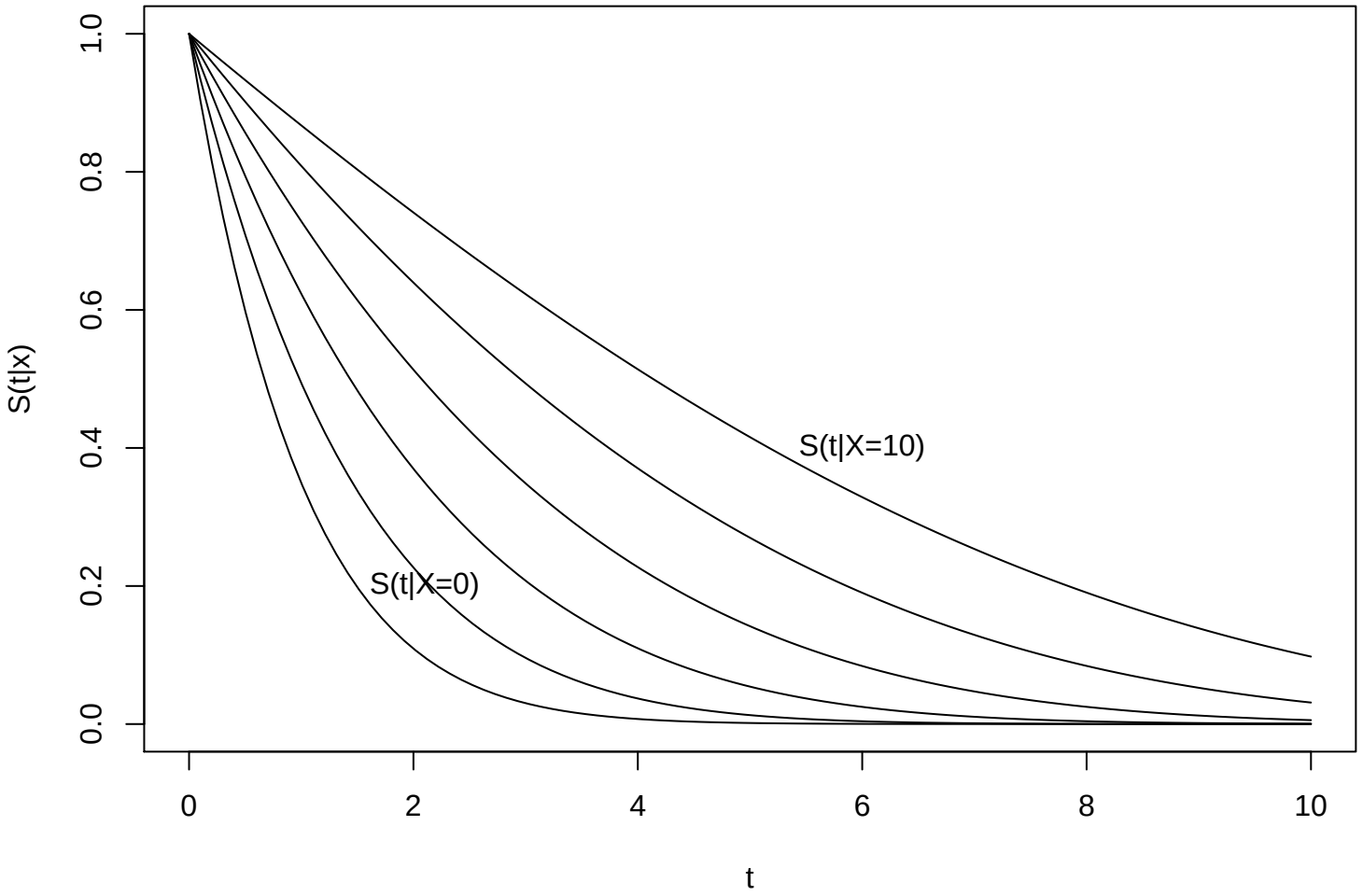

log hazards
log(h(t|x))
1.0
0.5
0.0
−0.5
−1.0
−1.5
−2.0
β
0
2
4
6
8
10
t


S(t|x)
1.0
0.8
0.6
0.4
0.2
0.0
S(t|X=1)
S(t|X=0)
0
2
4
6
8
10
t


Risk difference
S(t|X=0)−S(t|X=1)
0.08
0.06
0.04
0.02
0.00
0
2
4
6
8
10
t

Continuous example, $h(t|X) = h(t)\exp(X\beta)$.




S(t|x)
1.0
0.8
0.6
0.4
0.2
0.0
S(t|X=10)
S(t|X=0)
0
2
4
6
8
10
t

Specific parametric functions

- Exponential
- Weibull

Exponential proportional hazards regression

The exponential survival regression model can be expressed as

$$h(t|X) = \lambda \exp(X\beta)$$

$$S(t|X) = \exp[-\lambda t \exp(X\beta)] = \exp(-\lambda t)^{\exp(X\beta)}.$$

The regression can also be written as

$$\log h(t|X) = \log(\lambda) + X\beta.$$

If we replace λ with $\lambda = \exp(\beta_0)$, then

$$h(t|X) = \exp(\beta_0 + X\beta).$$

Therefore, we can think of λ as a transformed intercept term.

Example

Recall the ovarian cancer data set. We will fit the model

$$h(t|rx) = \lambda \exp(\beta rx),$$

where $rx = 1, 2$ is a treatment group indicator.

```
> se=survreg(Surv(futime, fustat)~rx, ovarian, dist='exponential')
> summary(se)

Call:
survreg(formula = Surv(futime, fustat) ~ rx, data = ovarian,
    dist = "exponential")
             Value Std. Error    z        p
(Intercept) 6.255      0.878 7.12 1.07e-12
rx          0.613      0.586 1.05 2.96e-01
```

We have to transform this output to interpret it in the proportional hazards setting.

$$\lambda = \exp(-(\text{Intercept})) = \exp(-6.255) = 0.00192$$

and

$$\beta = -\text{coefficient for } rx = -.613.$$

Therefore.

$$hr(rx = 2 : rx = 1) = \exp(-\beta) = \exp(-0.613) = 0.54$$

$$h(t|rx = 2) = \lambda \exp(2\beta) = 0.000564$$
$$h(t|rx = 1) = \lambda \exp(\beta) = 0.00104$$

Weibull example

The PH regression model for a Weibull distribution is defined as

$$h(t|X) = \alpha\gamma t^{\gamma-1}\exp(X\beta).$$

For the ovarian example, this becomes

$$h(t|X) = \alpha\gamma t^{\gamma-1}\exp(rx \times \beta).$$

```
> sg=survreg(Surv(futime, fustat)~rx  , ovarian, dist='weibull')
> summary(sg)

Call:
survreg(formula = Surv(futime, fustat) ~ rx, data = ovarian,
    dist = "weibull")
             Value Std. Error      z        p
(Intercept)  6.265      0.778  8.050 8.31e-16
rx           0.559      0.529  1.057 2.91e-01
```

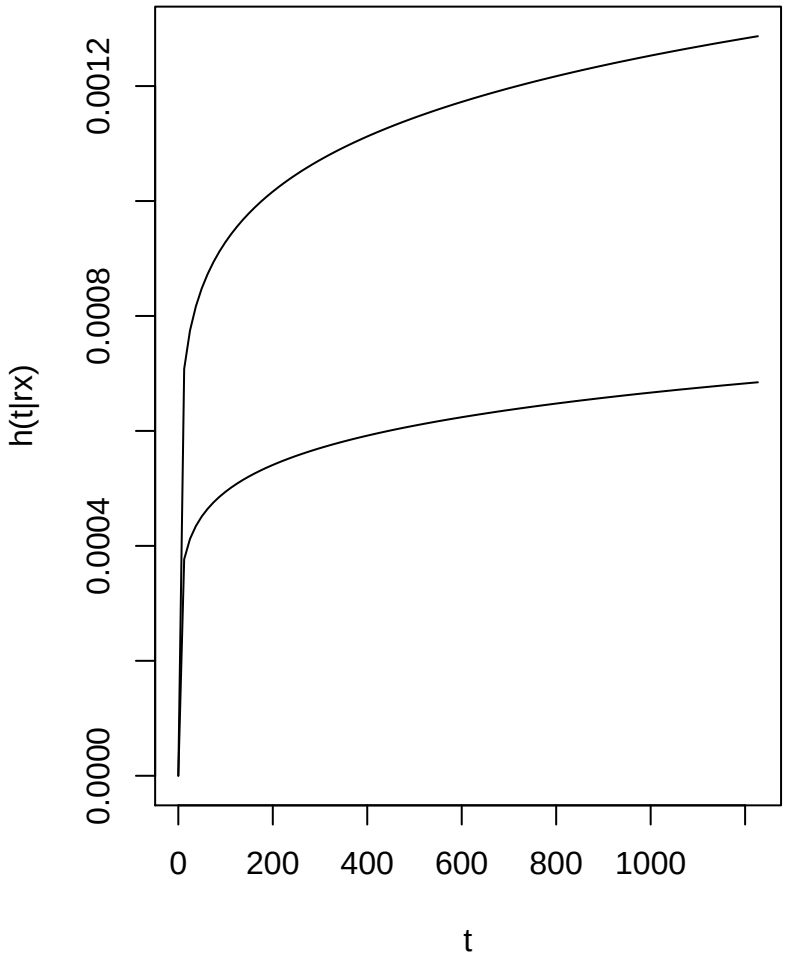
```
Log(scale)  -0.121      0.251 -0.483 6.29e-01

Scale= 0.886
```

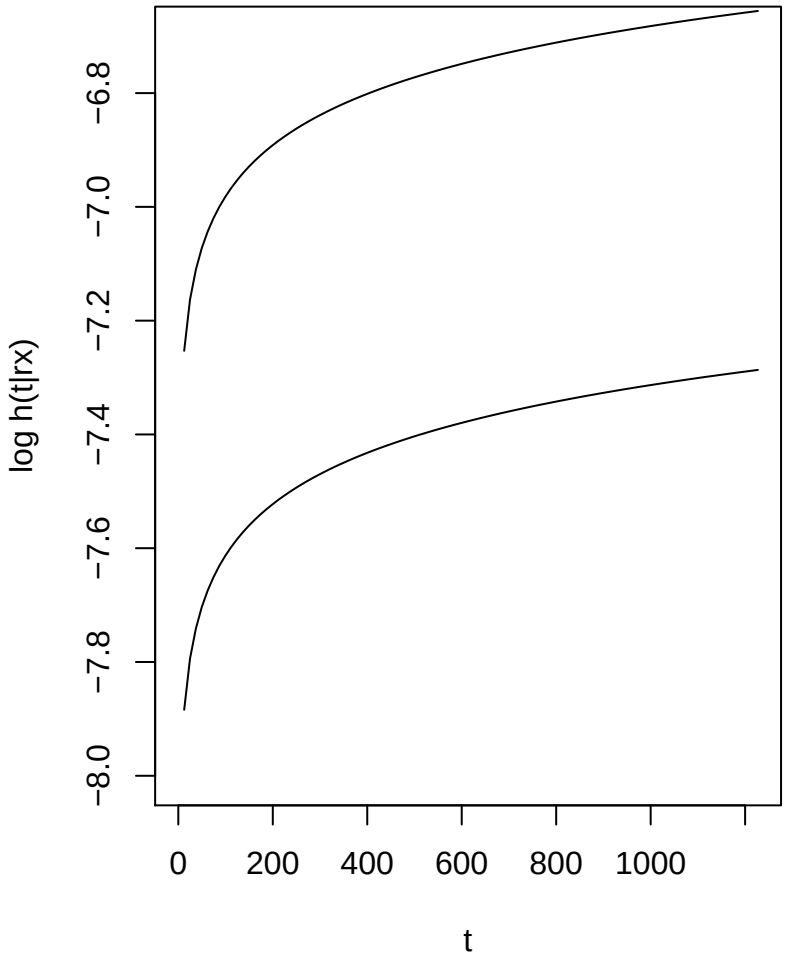
$$\gamma = 1/\text{Scale} = 1/0.886 = 1.13$$

$$\alpha = \exp(-(\text{Intercept})\gamma) = \exp(-6.265/0.886) = 0.000849$$

$$\beta = -\text{coefficient for } rx \times \gamma = -0.559/0.886 = -0.631$$

$$\begin{aligned} h(t|rx) &= \alpha\gamma t^{\gamma-1}\exp(rx\beta) \\ &= 0.000849 \times 1.13 t^{0.13}\exp(-0.631 rx) \end{aligned}$$


h(t|rx)
0.0000
0.0004
0.0008
0.0012
0
200
400
600
800
1000
t


log h(t|rx)
−8.0
−7.8
−7.6
−7.4
−7.2
−7.0
−6.8
0
200
400
600
800
1000
t

Accelerated failure time models

The *accelerated failure time (AFT) model* specifies that predictors act multiplicatively on the failure time (additively on the log of the failure time). The predictor alters the rate at which a subject proceeds along the time axis.

The model is

$$S(t|X) = \psi((log(t) - X\beta)/\sigma),$$

where ψ is any standard survival distribution and σ is called the scale parameter.

We can also write this relationship as

$$log(T) = X\beta + \sigma\epsilon,$$

where ϵ is a random variable from the ψ distribution.

Assumptions:

- The true form of ψ is correctly specified.
- Each X_j affects $\log(T)$ linearly (assuming no interactions).
- σ is a constant, independent of X.

The exponential and Weibull distributions are the only two distributions that can be used to describe both PH and AFT models.

These models can be fit in R using the $survreg()$ function.

Testing in parametric models

- As in logistic regression, parameter estimates in parametric survival models are obtained using maximum likelihood estimation.

- Therefore, we can use the same procedures for testing and constructing confidence intervals in parametric survival analysis as we did for logistic regression.

Using the ovarian data set, we fit the following Weibull regression model with age and treatment and predictors.

```
> sw2=survreg(Surv(futime, fustat)~rx+age  , ovarian, dist='weibull')
> summary(sw2)

Call:
```

```
survreg(formula = Surv(futime, fustat) ~ rx + age, data = ovarian,
    dist = "weibull")
              Value Std. Error     z        p
(Intercept) 10.4626     1.4427  7.25 4.10e-13
rx           0.5673     0.3403  1.67 9.55e-02
age         -0.0791     0.0198 -4.00 6.41e-05
Log(scale)  -0.5967     0.2352 -2.54 1.12e-02

Scale= 0.551

Weibull distribution
Loglik(model)= -88.8   Loglik(intercept only)= -98
        Chisq= 18.38 on 2 degrees of freedom, p= 1e-04
```

The column labeled z is the Wald statistic $(\hat{\beta}_j / \hat{se}(\hat{\beta}_j))$ for testing $H_0 : \beta_j = 0$.

Given the models fit in this lecture, how could we construct a likelihood ratio test for testing $\beta_{age} = 0$?

How could we construct a confidence interval for the hazard ratio?

Cox proportional hazards regression model

$h(t|X) = h(t)\exp(X\beta)$ is the proportional hazards regression model.

The Cox PH model

- is a semiparametric model
- makes no assumptions about the form of $h(t)$ (non-parametric part of model)
- assumes parametric form for the effect of the predictors on the hazard

In most situations, we are more interested in the parameter estimates than the shape of the hazard. The Cox PH model is well-suited to this goal.

Brief overview of estimation of β

Parameter estimates in the Cox PH model are obtained by maximizing the partial likelihood as opposed to the likelihood. The partial likelihood is given by

$$L(\beta) = \prod_{Y_i \text{ uncensored}} \frac{\exp(X_i\beta)}{\sum_{Y_j \geq Y_i} \exp(X_j\beta)}$$

The log partial likelihood is given by

$$l(\beta) = \log L(\beta) = \sum_{Y_i \text{ uncensored}} \{X_i\beta - \log[\sum_{Y_j \geq Y_i} \exp(X_j\beta)\}]$$

Cox and others have shown that this partial log-likelihood can be treated as an ordinary log-likelihood to derive valid (partial) MLEs of β.

The partial likelihood is valid when there are no ties in the data set. That is no two subjects have the same event time. If there are ties in the data set, the true partial log-likelihood function involves permutations and can be time-consuming to compute. In this case, either the Breslow or Efron approximations to the partial log-likelihood can be used.

Model assumptions and interpretations of parameters

- Same model assumptions as parametric model - except no assumption on the shape of the underlying hazard.

- Parameter estimates are interpreted the same way as in parametric models, except no shape parameter is estimated because we are not making assumptions about the shape of the hazard.

Example

$$h(t|rx, age) = h(t)\exp(\beta_1 \times rx + \beta_2 \times age)$$

```
> cph1=coxph(Surv(futime, fustat)~rx+age  , ovarian)
> summary(cph1)
Call:
coxph(formula = Surv(futime, fustat) ~ rx + age, data = ovarian)

  n= 26

       coef exp(coef) se(coef)     z      p
rx  -0.804     0.448   0.6320 -1.27 0.2000
age  0.147     1.159   0.0461  3.19 0.0014

    exp(coef) exp(-coef) lower .95 upper .95
rx      0.448      2.234     0.130      1.54
age     1.159      0.863     1.059      1.27
```

```
Rsquare= 0.457   (max possible= 0.932 )
Likelihood ratio test= 15.9  on 2 df,   p=0.000355
Wald test            = 13.5  on 2 df,   p=0.00119
Score (logrank) test = 18.6  on 2 df,   p=9.34e-05
```