STA2201H Methods of Applied Statistics II

Monica Alexander

Week 3: Survival Analysis

Overview

- Notes
- Definitions
- Constant hazards
- Kaplan Meier
- Parametric
- Semi-parametric
  - Piece-wise constant hazards
  - Cox proportional hazards

A note

$$y \sim \text{Poisson}(\lambda E)$$

We are interested in studying patterns in the rate, λ

Model (in R)

$$\log y = E(\lambda|\mathbf{X}) + \log E + \varepsilon_i$$

$$\log y = \mathbf{X}\beta + \log E + \varepsilon_i$$

Another note

In practice it may not be entirely clear what GLM to fit.

Example: you are microdata for multiple annual surveys in Kenya, and asked to study patterns in internal migration. You have info on things like

- Year, age, gender, county of residence, county of residence 1 year ago
- Income, education, marital status, number of kids
- Urban/rural, pop density of county, main industry of county
- ....

Another note ctd

- What's the likelihood?
- What sorts of covariates should you include in your model?
- How should these covariates be included? (e.g. age/year)

Survival Analysis

Some analysis from Duty/Shame

Some analysis from Duty/Shame

Some analysis from Duty/Shame

Recommended reading

- Hosmer et al (2008) 'Applied Survival Analysis: Regression Modeling of Time-to-Event Data'
- Dobson chapter 10
- German Rodriguez's course is a nice overview: https://data.princeton.edu/pop509/

Introduction

- Interested in the **waiting time** to an event / outcome
- Terminology is all around survival and death, but can be used to study any sort of waiting time
  - time to first birth
  - time to leaving home
  - time to finishing PhD :)
- Increasing amount of information considered (not just looking at end outcome)

Goals:

- Analyse waiting times wrt covariates
- Adjust for potential censoring or truncation

Survival analysis is a suite of methods to do this including parametric, semi-parametric and non-parametric methods.

Definitions

Let T be a non-negative random variable representing the waiting time to an event of interest.

- Assume T is continuous
- Define the pdf of T as $f(t)$
- cdf is $P(T < t) = F(t)$
- **Survival function** $P(T \geq t) = 1 - F(t) = S(t)$

Definitions

The **hazard rate** $\lambda(t)$ is the instantaneous rate of occurrence

$$\lambda(t) = \lim_{dt \to 0} \frac{Pr(t \leq T < t + dt | T \geq t)}{dt}$$

which is

$$\lambda(t) = \frac{f(t)}{S(t)} = -\frac{d}{dt} \log(S(t))$$

Definitions

$$\lambda(t) = \frac{f(t)}{S(t)} = -\frac{d}{dt}\log(S(t))$$

implies

$$S(t) = \exp\left(-\int_0^t \lambda(x)dx\right) = \exp\left(-\Lambda(t)\right)$$

where $\Lambda(t)$ is the cumulative hazard $=$ the sum of risks one faces up to time t.

Given either the hazard rate or survival function, you can get everything else.

Characteristics shapes for deaths in human populations

Example: constant hazards

$$S(t) = \exp\left(-\int_0^t \lambda(x)dx\right) = \exp\left(-\Lambda(t)\right)$$

- The simplest case is if $\lambda(t) = \lambda$ for all t
- Constant hazard of dying/event occurring

This implies

$$S(t) = \exp(-\lambda t)$$

and

$$f(t) = \lambda \exp(-\lambda t)$$

What is this distribution?

Likelihood

Individuals i, $i = 1, \ldots, n$

- By any particular time t_i, i is either alive or dead
- If they are alive, they are **censored**
- Contribution to likelihood if died: $f(t_i) = \lambda(t_i)S(t_i)$
- Contribution to likelihood if alive: $S(t_i)$

Likelihood is then

$$L = \prod_i L_i = \prod_i \lambda(t_i)^{d_i} S(t_i)$$

d_i is indicator of whether or not individual died.

and LL is

$$\log L = \sum_i d_i \log(\lambda(t_i)) - \Lambda(t_i)$$

Example: constant hazards

If $\lambda(t) = \lambda$ then

$$\log L = \sum_i d_i \log \lambda - \sum_i t_i \lambda$$

So MLE for λ is ?

Example: constant hazards

$$\hat{\lambda} = \frac{\sum d_i}{\sum t_i}$$

- if nothing is censored this is just Exponential

Example: constant hazards

Instead of looking at waiting times, look at deaths $\sum_i d_i = D$ and assume

$$D \sim \text{Poisson}(\lambda E)$$

where $E = \sum_i t_i$.

What's the MLE?

$$\log L = \sum_i d_i \log \lambda - \sum_i t_i \lambda + \text{constant}$$

- Waiting times Exponential == deaths are Poisson
- Helps to use GLM!

Where we are at so far

- Likelihood is straightforward to write down for i individuals with observed times t_i
  - Contribution to likelihood if died: $f(t_i) = \lambda(t_i)S(t_i)$
  - Contribution to likelihood if alive: $S(t_i)$
  - Indicator variable $d_i = 1$ if i died and 0 otherwise.

Likelihood is then

$$L = \prod_i L_i = \prod_i \lambda(t_i)^{d_i} S(t_i)$$

- If we assume the hazard is constant over time, $\lambda(t) = \lambda$:
  - this is equivalent to saying the waiting times t_i are Exponential
  - this is also equivalent to saying that the deaths are Poisson distributed with rate λE where E is total exposure time ($E = \sum_i t_i$).

Where we are going

A primary goal of survival analysis: to estimate the survival function $S(t)$.

Can do this in three ways:

1. Non-parametric (Kaplan-Meier)
2. Parametric (choose your weapon)
3. Semi-parametric (Piecewise constant hazards)

Kaplan-Meier Estimator

First: discrete hazard and survival times

Imagine we have a set of ordered observations of survival times $t_1 < t_2 < \ldots t_n$.

- pmf $f(t_i) = f_i = Pr(T = t_i)$
- survival function $S(t_i) = S_i = P(T \geq t_i) = 1 - \sum_i^{\infty} f(t_i)$
- hazard function $\lambda(t_i) = \lambda_i = Pr(T = t_i | T \geq t_i) = f(t_i)/S(t_i)$

Can also write discrete Survival function as

$$S_i = \prod_{i:t_i \leq t} (1 - \lambda_i)$$

Intuition: product of the conditional probabilities of survival.

Kaplan-Meier Estimator

One way of estimating $S(t)$ is to not assume any functional form at all and estimate directly from the data.

The K-M estimator is:

$$\hat{S}(t) = \prod_{i:t_i \leq t} \left(1 - \frac{d_i}{n_i}\right)$$

- t_i is a time where at least one event happened (in between times, $S(t)$ is constant)
- d_i is the number of events
- n_i is the number or individuals still at risk at time t_i (i.e. not dead or censored)

Intuition: our best guess at probability of death at time t_i is just deaths divided by exposure.

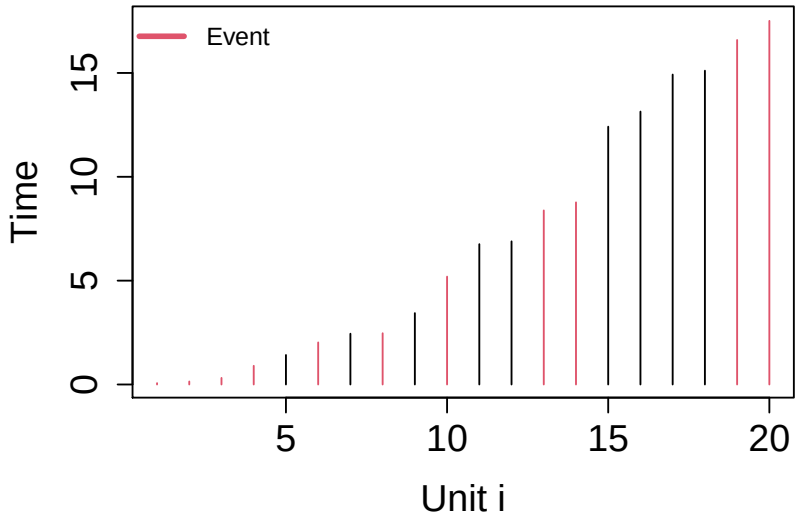
Kaplan-Meier Estimator

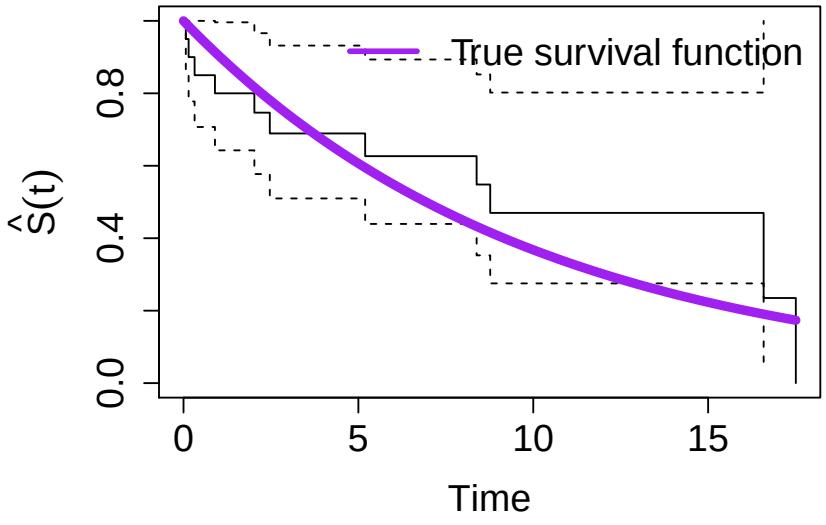
Variance is:

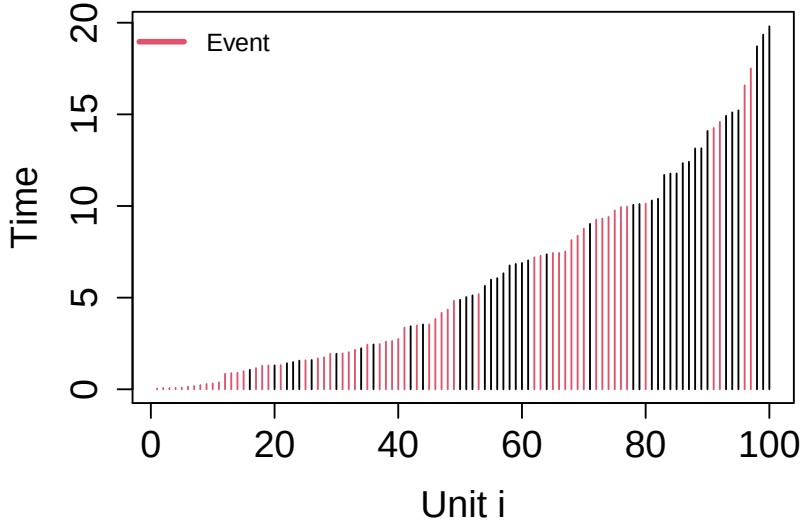
$$Var(\hat{S}(t)) = (\hat{S}(t))^2 \sum_{i:t_i \leq t} \frac{d_i}{n_i(n_i - d_i)}$$

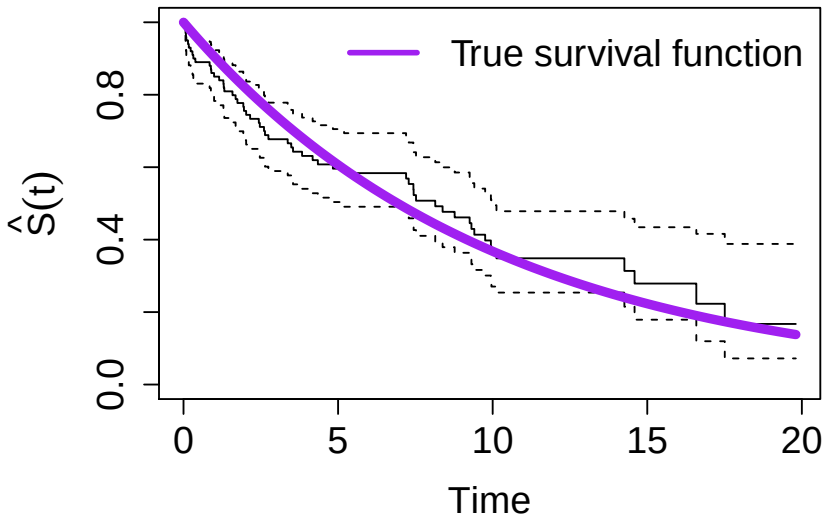
(obtained using Delta method twice, and assuming deaths are binomial)

Kaplan-Meier Estimator









Kaplan-Meier Estimator

If you have a set of survival times, t and an indication of whether or not they are censored:

How to calculate:

1. Order your observations
2. Calculate n_i i.e. number still at risk of event at each time point (i.e. those who haven't died or become censored)
3. Calculate d_i/n_i for each time point
4. Do the cumulative product of $1 - d_i/n_i$

OR

Use the `survival` package

(More in lab)

Kaplan-Meier Estimator

Why is it good?

Why is it bad?

Parametric survival functions

Parametric estimation

Assume the survival times follow a specified parametric form.

- Estimate $S(t)$ with a smooth function
- Easily convert to $\lambda(t)$ etc
- Estimate useful quantities like mean, quantiles etc
- Extrapolate past the last observation

But only good if the parametric form is correct.

Common distributions

- Exponential (hazards: constant)

$$f(t) = \lambda \exp(-\lambda t)$$

- Weibull (hazards: monotonically increasing or decreasing)

$$f(t) = \frac{a}{b}\left(\frac{t}{b}\right)^{(a-1)} \exp\left(-\frac{t}{b}^{a}\right)$$

- Gamma (hazards: monotonically increasing or decreasing)

$$f(t) = \frac{\beta^{\alpha}}{\Gamma(\alpha)} t^{\alpha-1} \exp(-\beta t)$$

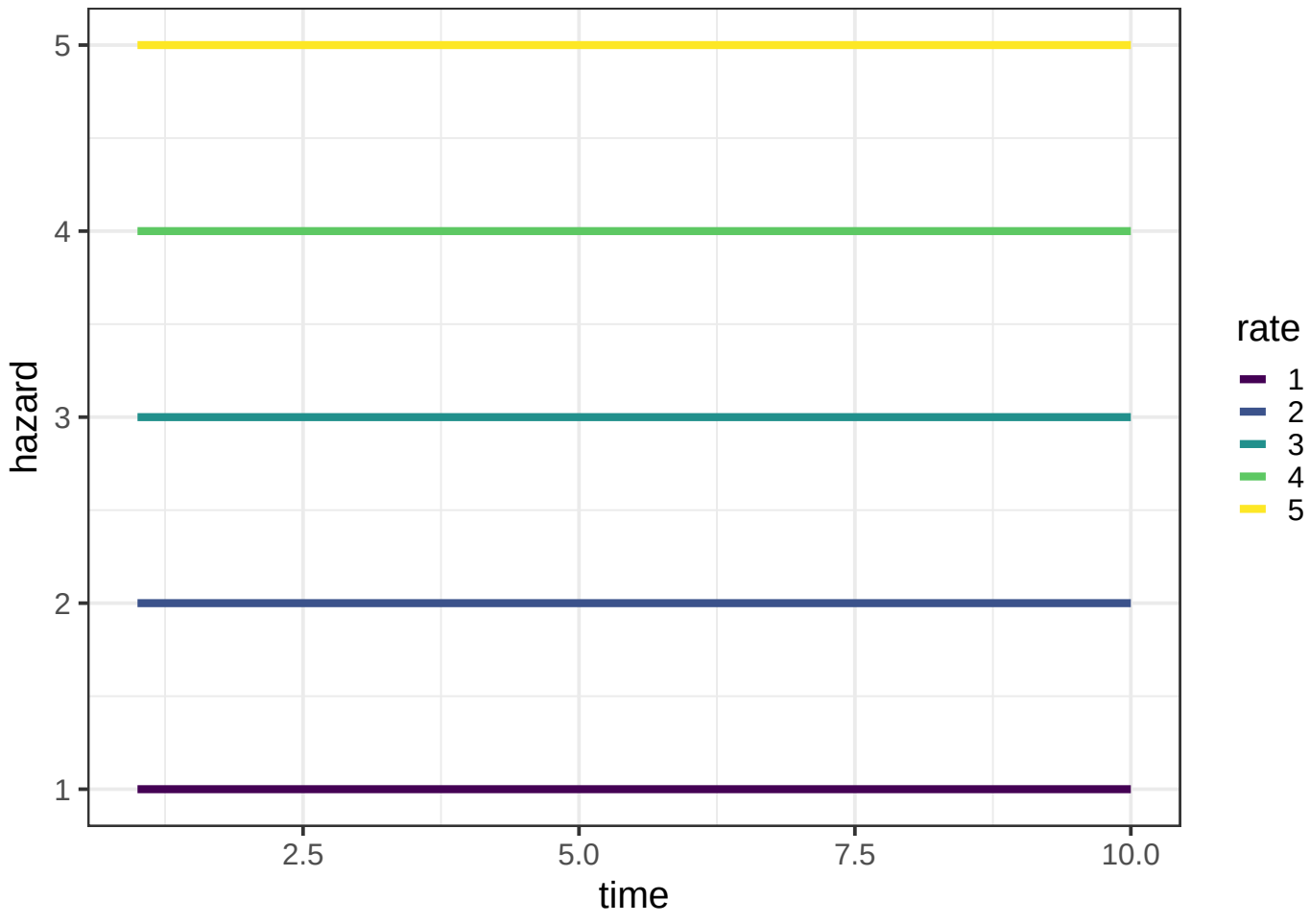
- Lognormal (hazards: arc-shaped and monotonically decreasing)

$$f(t) = \frac{1}{t\sigma\sqrt{2\pi}} e^{-\frac{(\ln t - \mu)^2}{2\sigma^2}}$$

Shapes of $\lambda(t)$

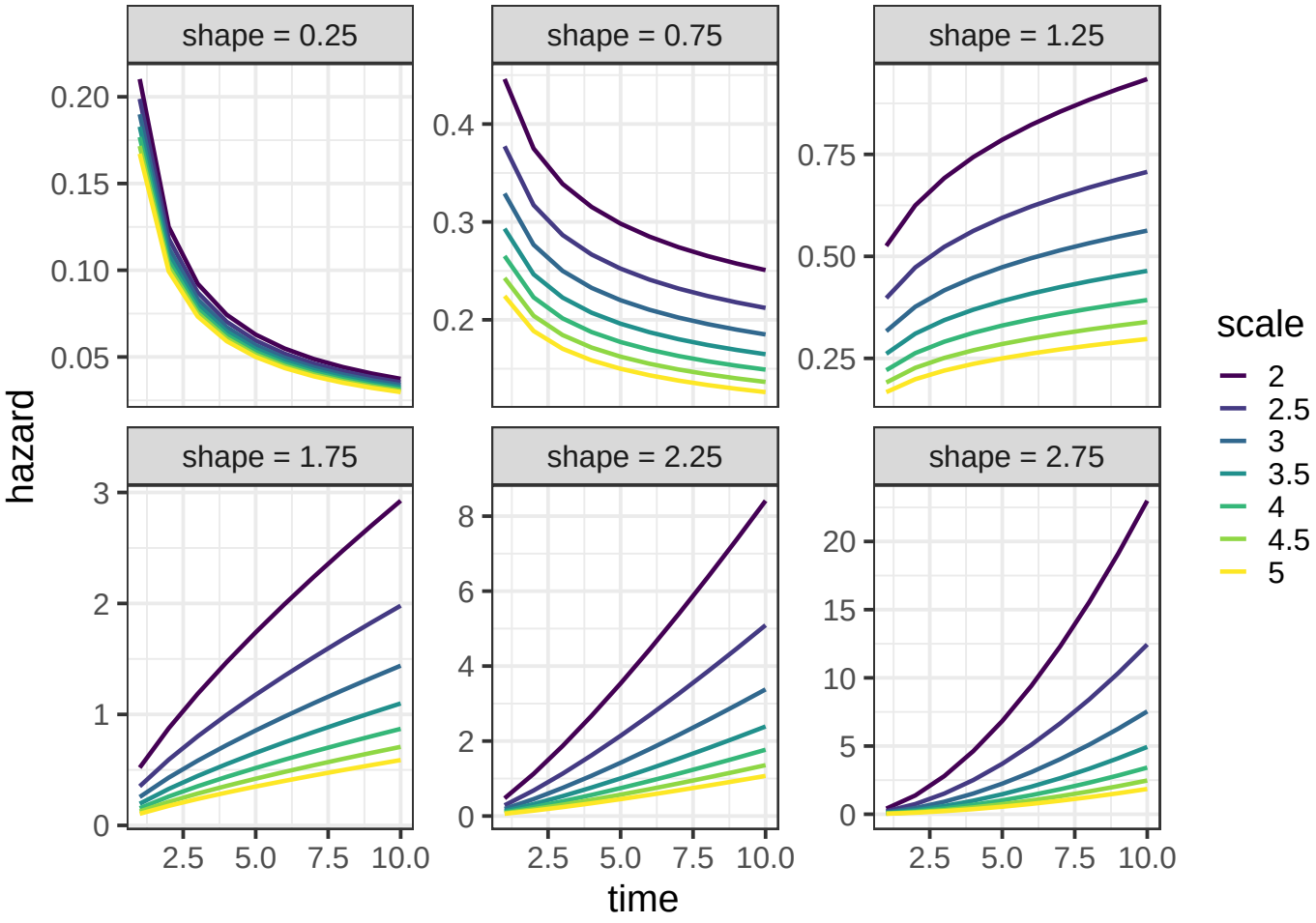
Exponential

$$f(t) = \lambda \exp(-\lambda t)$$



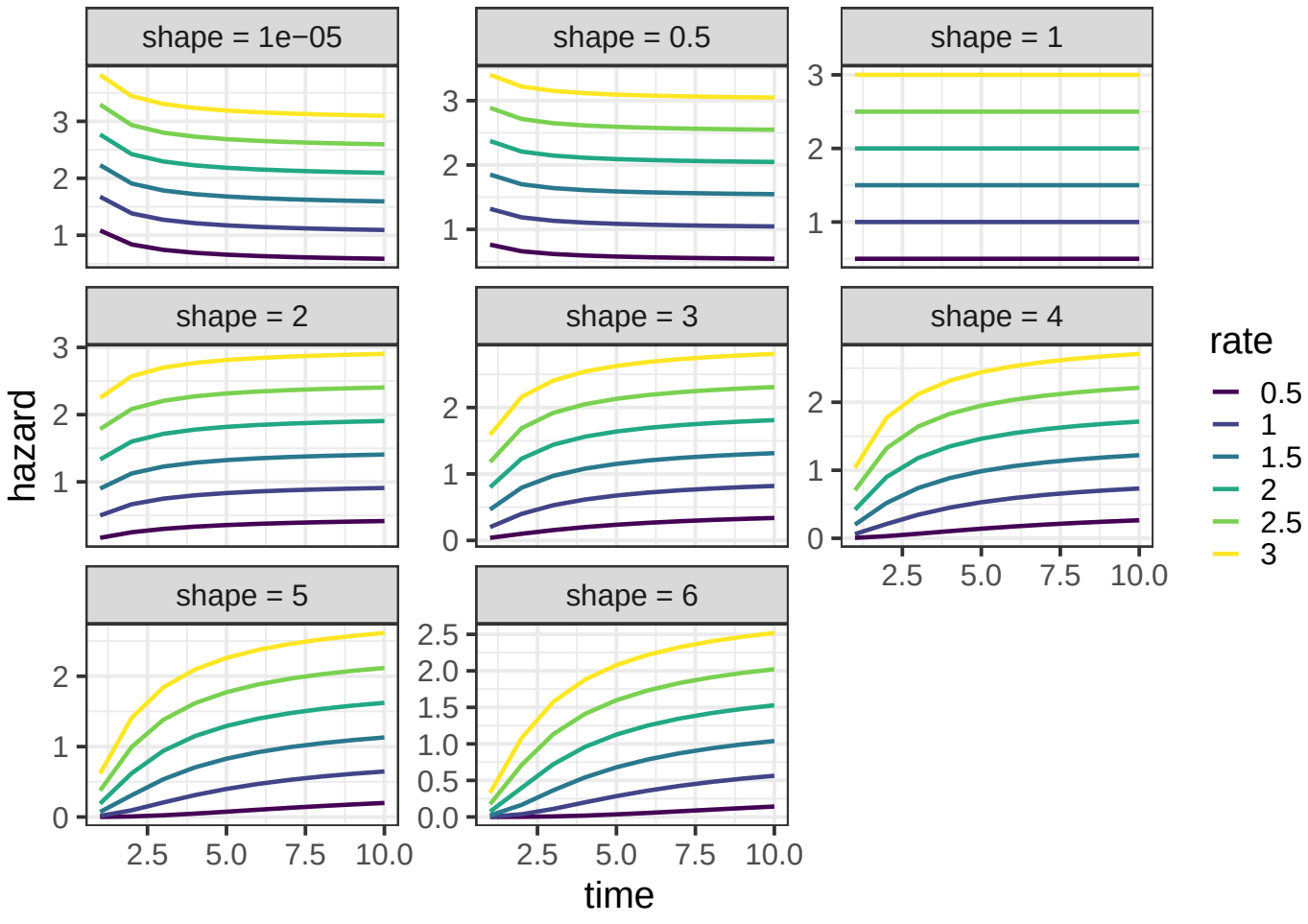
Weibull

$$f(t) = \frac{a}{b}\left(\frac{t}{b}\right)^{(a-1)} \exp\left(-\frac{t}{b}^{a}\right)$$



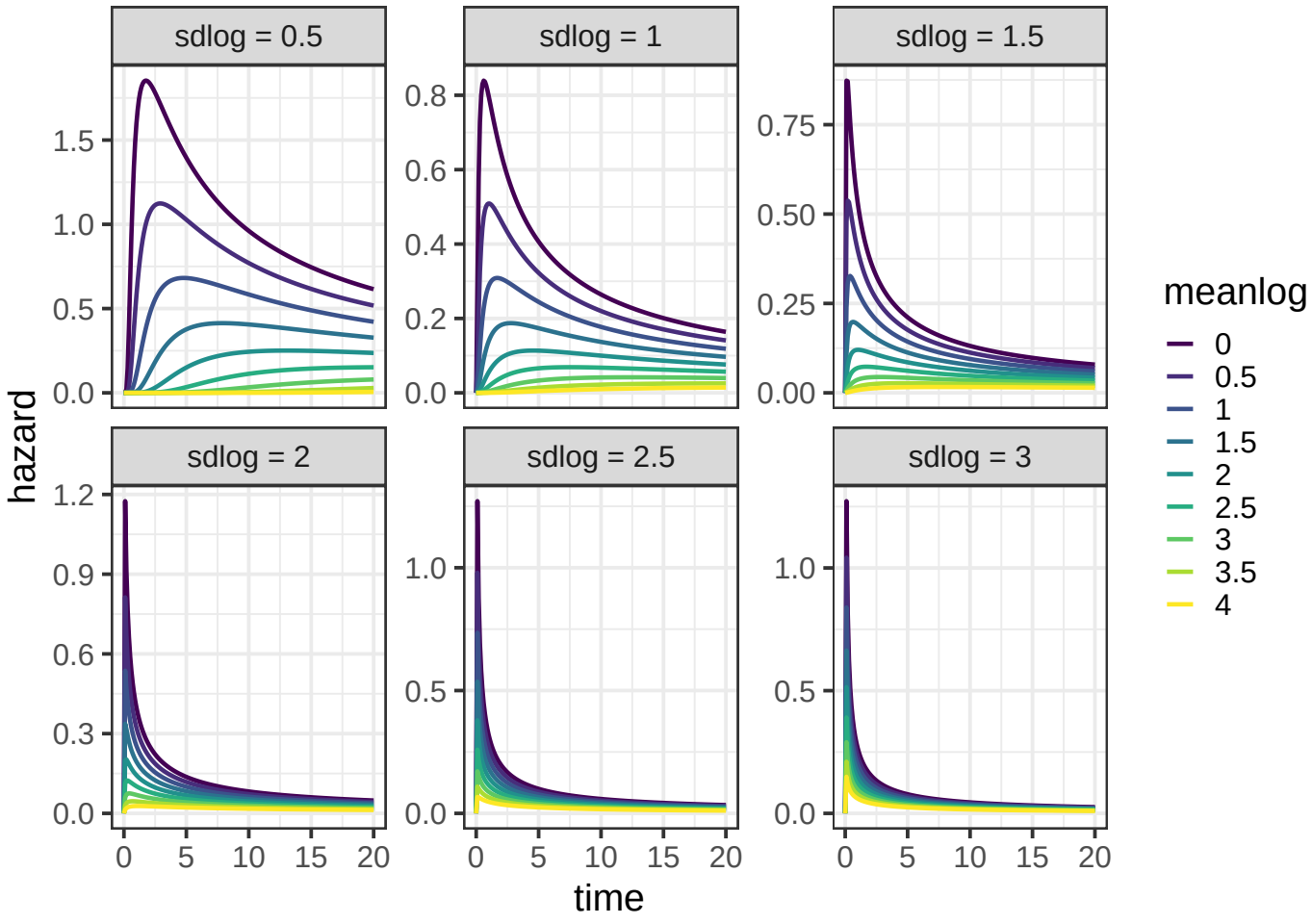
Gamma

$$f(t) = \frac{\beta^{\alpha}}{\Gamma(\alpha)} t^{\alpha-1} \exp(-\beta t)$$



Lognormal

$$f(t) = \frac{1}{t\sigma\sqrt{2\pi}} e^{-\frac{(\ln t - \mu)^2}{2\sigma^2}}$$



Parametric survival models in R

In addition to built-in functions, the library `flexsurv` is super useful here.

- density, distribution, quantile, random generation: most in base `stats` e.g.:

```
rweibull(1, shape = 1.25, scale = 2)
```

```
## [1] 2.103756
```

- hazard function: built-in functions in `flexsurv` e.g.

```
hweibull(10, shape = 1.25, scale = 2)
```

```
## [1] 0.934593
```

Parametric survival models in R

Fitting to data to get parameter estimates: fit intercept only model using `flexsurvreg`

E.g. Fit a Weibull

```
flexsurvreg(Surv(time, status) ~ 1, data = d, dist = "Weibull")
```

```
## Call:
## flexsurvreg(formula = Surv(time, status) ~ 1, data = d, dist = "Weibull")
##
## Estimates:
##        est       L95%      U95%      se
## shape    1.3168    1.1652    1.4882    0.0822
## scale  417.7587  372.0394  469.0963   24.7045
##
## N = 228,  Events: 165,  Censored: 63
## Total time at risk: 69593
## Log-likelihood = -1153.851, df = 2
## AIC = 2311.702
```

Parametric survival models

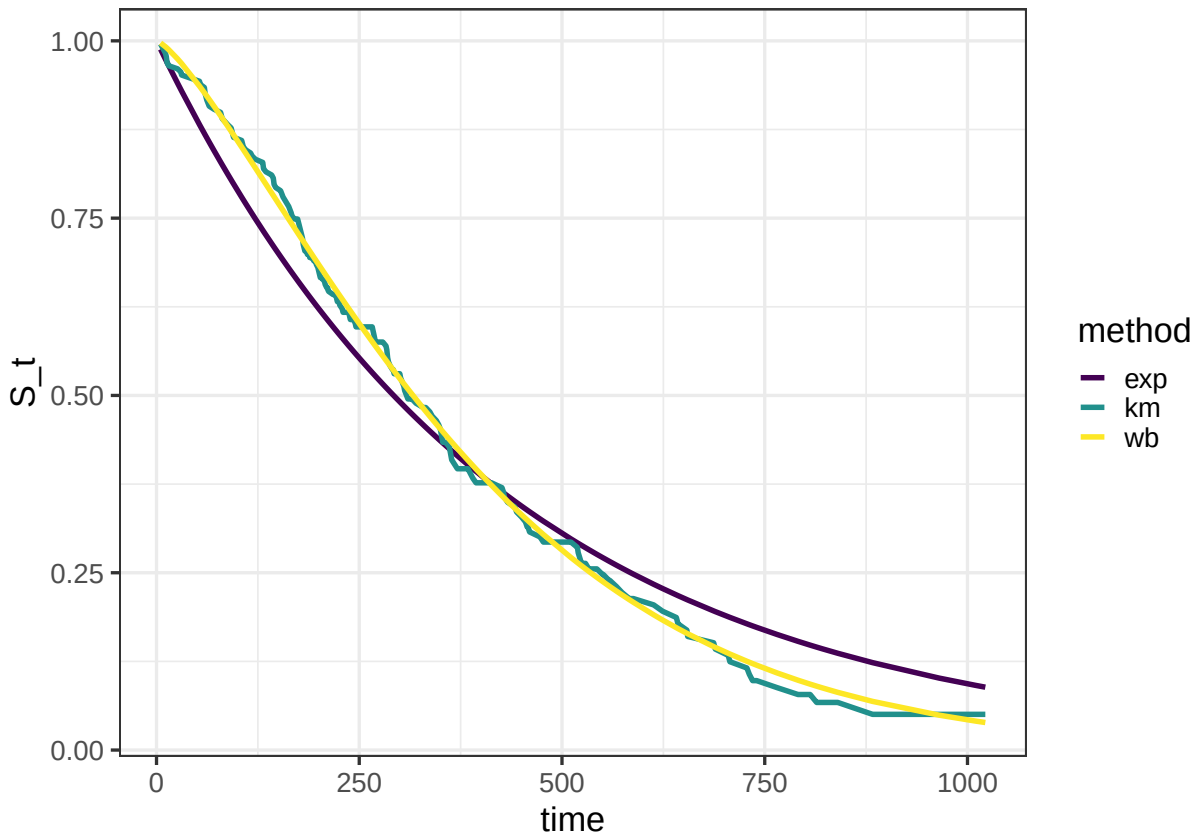
- How do decide on parametric form?
- How to assess whether this is reasonable?

Data exploration <-> Assessing fit ad infinitum

```
fit_km <- survfit(Surv(time, status)~1, data = d)
fit_w <- flexsurvreg(Surv(time, status) ~ 1,
                     data = d, dist = "Weibull")
fit_e <- flexsurvreg(Surv(time, status) ~ 1,
                     data = d, dist = "Exponential")

res_df <- tibble(time = sort(unique(d$time)), km = fit_km$surv,
       wb = 1- pweibull(sort(unique(d$time)),
                        shape = exp(fit_w$coefficients[1]),
                        scale = exp(fit_w$coefficients[2])),
       exp =  1- pexp(sort(unique(d$time)),
                      rate = exp(fit_e$coefficients[1])))
```

Data exploration <-> Assessing fit ad infinitum

```
res_df %>%
  pivot_longer(-time, values_to = "S_t",
               names_to = "method") %>%
  ggplot(aes(time, S_t, color = method)) +
  geom_line(lwd = 1.5) +
  theme_bw(base_size = 20) + scale_color_viridis_d()
```



Semi-parametric estimation

Semi-parametric survival

- Make milder assumptions about underlying hazards

Piecewise Constant Hazards:

- Divide time into reasonably small intervals and assume that the baseline hazard is constant in each interval

Back to Exponential / Poisson equivalency

Recall that if we assume constant hazards, then survival times t_i are

$$t_i \sim \text{Exp}(\lambda)$$

or equivalently, consider total events $D = \sum_i d_i$ (where $d_i = 1$ if event occurred) and total exposure $E = \sum_i t_i$ then total events ("deaths") are

$$D \sim \text{Poisson}(\lambda E)$$

Back to Exponential / Poisson equivalency

Sanity check in R: simulated exponential waiting times with rate 0.1. The `survobject` has times `t.i` and whether or not the event happened `d.i`

```
fitE <- flexsurvreg(survobject  ~ 1, dist = "Exponential")
lambdahat <- exp(coef(fitE))

fitP <- glm(D ~ offset(log(E)), family  = "poisson")
lambdahatP <- exp(coef(fitP))

list("poisson" = lambdahatP[[1]], "exp" =  lambdahat, "data" =  D/E)
```

```
## $poisson
## [1] 0.09484268
##
## $exp
## [1] 0.09484268
##
## $data
## [1] 0.09484268
```

PCH set-up

- Suppose we have partitioned duration time into C intervals, defined by cut points (times) $0 = \tau_0, \tau_1, \ldots, \tau_C = \infty$ such that the kth interval is given by $[\tau_{k-1}, \tau_k)$.
- If we assume that the hazard is constant in each interval, we get

$$\lambda(t) = \lambda_k \text{ for } t \text{ in } [\tau_{k-1}, \tau_k)$$

- The parameter vector θ to define the hazard $\lambda(t|C)$ is

$$\theta = (\lambda_1, \lambda_2, \ldots, \lambda_C)$$

PCH set-up: waiting times (individuals)

- Focus on what happens in the kth interval is given by $[\tau_{k-1}, \tau_k)$:
  - At the start of the interval, we are left with a subset of individuals, those with $t_i \geq \tau_{k-1}$
  - The hazard is constant, thus for event time T_i for individual i who is still around, we get

$$(T_i - \tau_{k-1}) \sim \text{Exp}(\lambda_k)$$

    up to censoring time τ_k
  - This is no different what we had before, we just restart counting events and exposure at each cut point.

PCH set-up: death counts (groups)

$$D_k \sim \text{Poisson}(\lambda_k E_k)$$

where

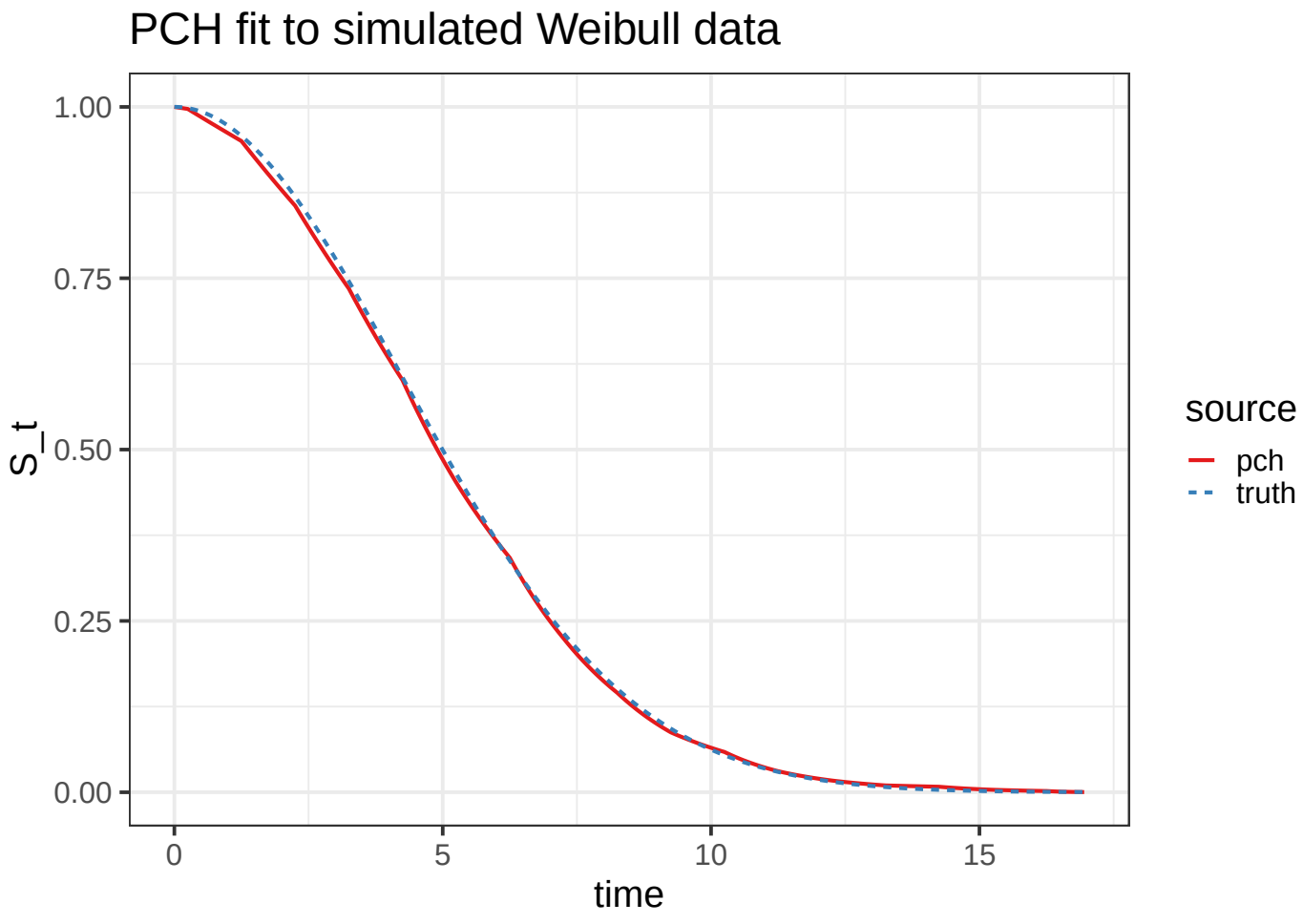
$$D_k = \sum_i d_i \cdot 1(\tau_{k-1} \leq t_i < \tau_k)$$

and

$$E_k = \sum_i 1 \cdot (t_i \geq \tau_{k-1}) \cdot (\min(t_i, \tau_k) - \tau_{k-1})$$

Simulation illustration

Neat: with appropriately chosen intervals and sufficient information, can provide close fits to survival functions that are in fact generated by continuous hazard functions.



Example: time to second birth

Data on birth intervals for married women with at least one birth, 19th northern Sweden. Part of the `eha` package in R. We will focus on women at parity 1 (women who have one child and the time until the second child).

The data look like:

id	next.ivl	event
1	22.348	0
2	1.837	1
3	2.051	1
4	1.782	0
5	1.629	1
6	1.730	1

Time to second birth

- define some cut-points
- `survSplit` is useful to get data in the right format (more in lab)
- Here's a sanity check, that the Poisson-estimated rates are just the same as the MLE

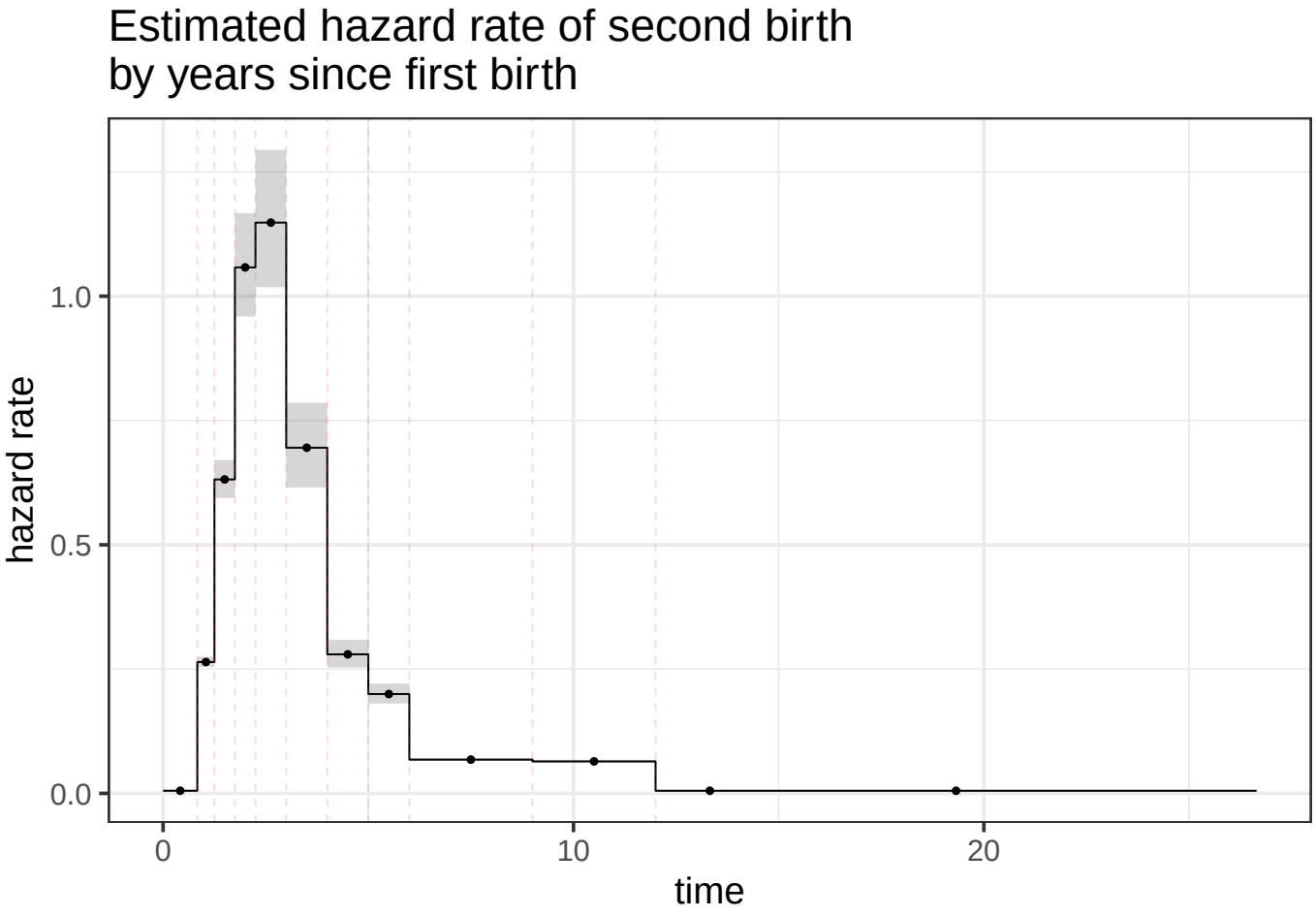
```
cutpoints <- c(10/12, 1.25, 1.75, 2.25, seq(3,5), seq(6, 12, by = 3))
C <- length(cutpoints) + 1

f12$enter <- 0  # everyone enters at time 0

f12_split <- survSplit(formula = Surv(time = next.ivl, event = event) ~ .,
                       data = f12, cut = cutpoints) %>% as_tibble() %>%
  mutate(interval = factor(tstart),
         interval_length = next.ivl - tstart)

E_k <- f12_split %>% group_by(interval) %>% summarise(E = sum(next.ivl-tstart)) %>% select(E) %>% pull()
D_k <- f12_split %>% group_by(interval) %>% summarise(E = sum(event)) %>% select(E) %>% pull()

intervals <- as.factor(1:length(D_k)) # number of intervals
fit_pois <- glm(D_k ~ offset(log(E_k))-1 + intervals, family = "poisson")
```

Sanity check

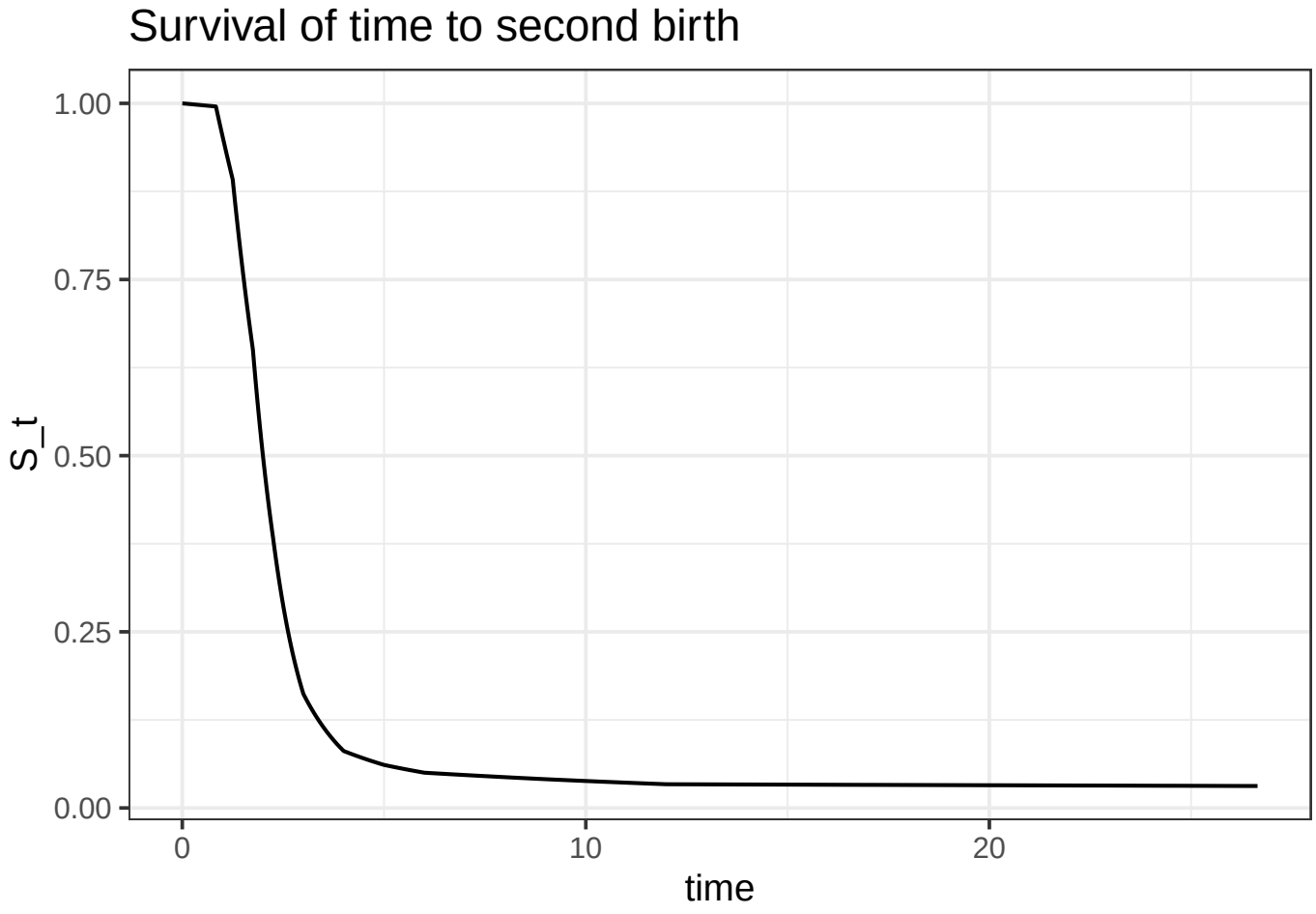
```
# compare:
round(data.frame(exp(coef(fit_pois)), D_k/E_k),4) %>%
  kable()
```

	exp.coef.fit_pois..	D_k.E_k
intervals1	0.0053	0.0053
intervals2	0.2642	0.2642
intervals3	0.6314	0.6314
intervals4	1.0580	1.0580
intervals5	1.1480	1.1480
intervals6	0.6953	0.6953
intervals7	0.2796	0.2796
intervals8	0.1998	0.1998
intervals9	0.0679	0.0679
intervals10	0.0643	0.0643
intervals11	0.0053	0.0053

Plot the hazards



Plot the implied survival curve



PCH models

- Super flexible, nice middle ground between KM and parametric
- Problem: where should the cut-points be?
  - substantive knowledge
  - data availability
  - standard tests e.g. LR test for subsetted models

Adding covariates

Adding covariates

- So far, have only looked at estimating one hazard function / survival function per study
- But want to see how different covariates are associated with survival
- Straight-forward to extend what we have already done, based on **proportional hazards assumption**
  - Most common covariate modeling strategy in survival analysis
  - The increase or reduction in risk in some group of interest is the same at all durations t

Proportional hazards in PCH models

- Let $\lambda_i(t)$ be the hazard for individual i, with covariates x_i.
  - e.g. for births example, say x_i=1 if woman is less than 30 years old
- Based on the proportional hazards assumption, if the covariate increases/decreases the hazard by the same factor $\exp(\beta)$ in each interval, we get

$$\lambda_i(t) = \lambda_k \cdot \exp(\beta x_i) \text{ for } t \text{ in } [\tau_{k-1}, \tau_k)$$

- In the age example, if younger women have shorter birth intervals, is $\beta > 0$ or < 0?
- Interpretation of coefficient (or $\exp(\beta)$)?

PCH regression example

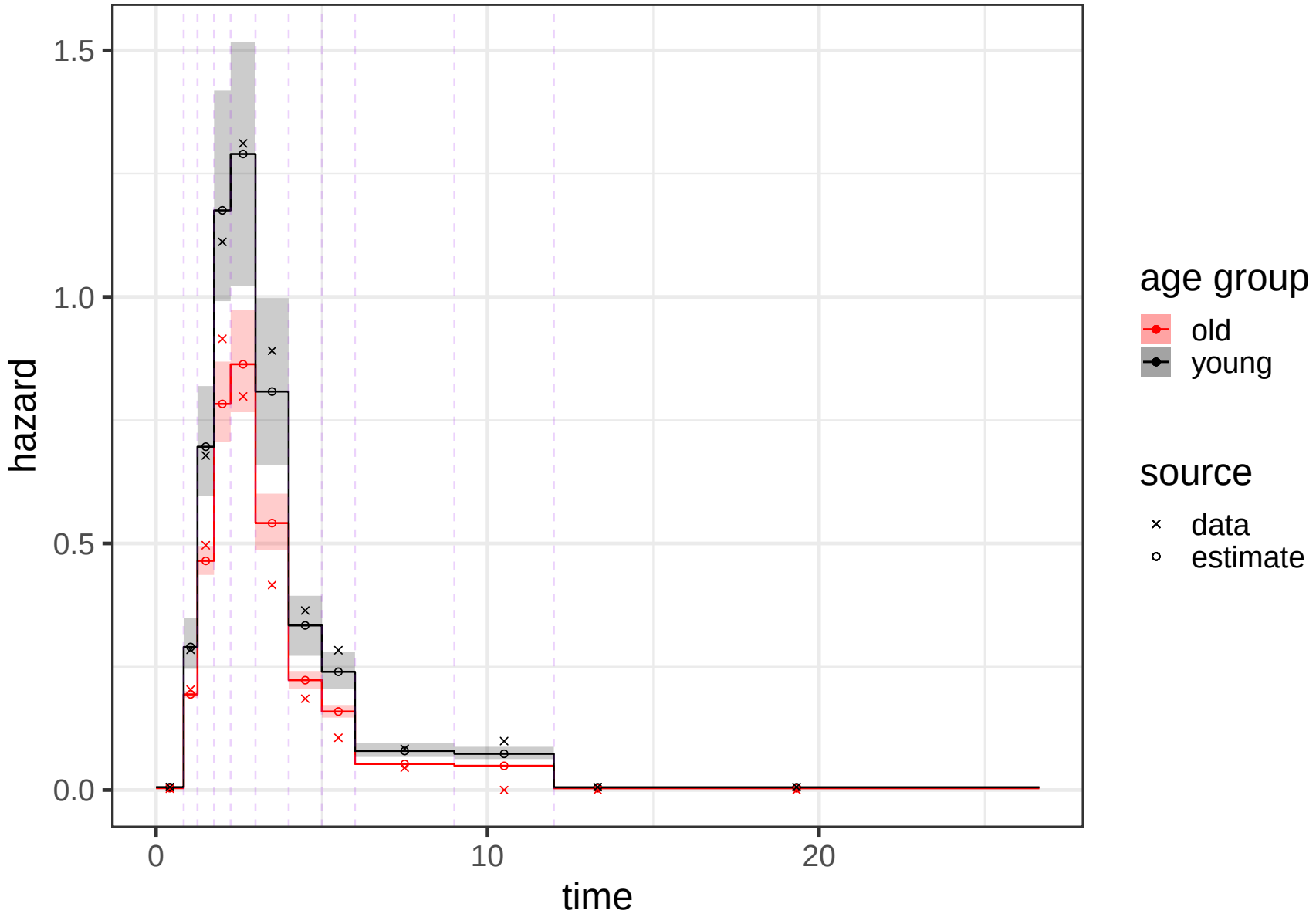
Back to our time to second birth dataset:

id	age	next.ivl	event	age_group
1	25	22.348	0	young
2	19	1.837	1	young
3	24	2.051	1	young
4	35	1.782	0	old
5	28	1.629	1	young
6	25	1.730	1	young

Run the regression

```
##
## Call:
## glm(formula = event ~ offset(off) + interval - 1 + age_group,
##     family = "poisson", data = f12_split)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.3861  -0.8130  -0.4020  -0.0806   4.1041
##
## Coefficients:
##                           Estimate Std. Error z value Pr(>|z|)
## interval0                 -5.54759    0.35688 -15.545  < 2e-16 ***
## interval0.833333333333333 -1.64042    0.08787 -18.669  < 2e-16 ***
## interval1.25              -0.76634    0.06812 -11.251  < 2e-16 ***
## interval1.75              -0.24466    0.06644  -3.683 0.000231 ***
## interval2.25              -0.14683    0.06922  -2.121 0.033887 *
## interval3                 -0.61378    0.09662  -6.353 2.12e-10 ***
## interval4                 -1.50168    0.18065  -8.312  < 2e-16 ***
## interval5                 -1.83794    0.25276  -7.271 3.56e-13 ***
## interval6                 -2.94110    0.30424  -9.667  < 2e-16 ***
## interval9                 -3.01649    0.38048  -7.928 2.23e-15 ***
## interval12                -5.59580    1.00147  -5.588 2.30e-08 ***
## age_groupyoung             0.39438    0.05965   6.611 3.80e-11 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for poisson family taken to be 1)
##
##     Null deviance: 10885.5  on 7625  degrees of freedom
## Residual deviance:  6314.6  on 7613  degrees of freedom
## AIC: 9652.6
##
## Number of Fisher Scoring iterations: 7
```

Plot the hazards by age group



Non-proportional hazards

What if instead:

- Births within a 3 year interval are more likely for older women,
- Births after a longer interval are less likely for older women.

The proportional hazard assumption is violated, and the model needs to be extended.

Non-proportional hazards can be modeled through the inclusion of interaction terms between the covariate and the duration time; for the PCH model, we can estimate interval-specific coefficients:

$$\lambda_i(t) = \lambda_k \cdot \exp(\beta_k x_i) \text{ for } t \text{ in } [\tau_{k-1}, \tau_k)$$

In our example, $\exp(\lambda_k + \beta_k)$ is the hazard for the younger women in interval k.

Fit in R

```r
fit2 <- glm(event ~ offset(off) + interval -1 + age_group + interval:age_group,
            family = "poisson", data = f12_split)
summary(fit2)
```

```
##
## Call:
## glm(formula = event ~ offset(off) + interval - 1 + age_group +
##     interval:age_group, family = "poisson", data = f12_split)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.4025  -0.8236  -0.4120  -0.0669   4.0879
##
## Coefficients:
##                                          Estimate Std. Error z value Pr(>|z|)
## interval0                                -5.92041    1.00000  -5.920 3.21e-09
## interval0.833333333333333                -1.59125    0.16667  -9.548  < 2e-16
## interval1.25                             -0.70039    0.10660  -6.570 5.02e-11
## interval1.75                             -0.08868    0.09535  -0.930   0.3523
## interval2.25                             -0.22534    0.11111  -2.028   0.0426
## interval3                                -0.87767    0.17678  -4.965 6.88e-07
## interval4                                -1.68640    0.31623  -5.333 9.67e-08
## interval5                                -2.24469    0.50000  -4.489 7.14e-06
## interval6                                -3.09522    0.57735  -5.361 8.27e-08
## interval9                               -16.14418  314.28305  -0.051   0.9590
## interval12                              -15.66239  309.40135  -0.051   0.9596
## age_groupyoung                            0.83393    1.06904   0.780   0.4353
## interval0.833333333333333:age_groupyoung -0.50009    1.08501  -0.461   0.6449
## interval1.25:age_groupyoung              -0.52160    1.07569  -0.485   0.6277
## interval1.75:age_groupyoung              -0.63931    1.07460  -0.595   0.5519
## interval2.25:age_groupyoung              -0.33756    1.07643  -0.314   0.7538
## interval3:age_groupyoung                 -0.07185    1.08826  -0.066   0.9474
## interval4:age_groupyoung                 -0.15807    1.13504  -0.139   0.8892
## interval5:age_groupyoung                  0.15029    1.21499   0.124   0.9016
## interval6:age_groupyoung                 -0.22092    1.26538  -0.175   0.8614
```

Try a reduced interaction

```r
f12_split <- f12_split %>%
  mutate(age_young_less_than_3 = ifelse(tstart<3&age_group=="young", 1, 0))

fit3 <- glm(event ~ offset(off) + interval -1 + age_group + factor(age_young_less_than_3),
            family = "poisson", data = f12_split)
summary(fit3)
```

```
##
## Call:
## glm(formula = event ~ offset(off) + interval - 1 + age_group +
##     factor(age_young_less_than_3), family = "poisson", data = f12_split)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.3748  -0.8249  -0.4130  -0.0828   4.1072
##
## Coefficients:
##                                 Estimate Std. Error z value Pr(>|z|)
## interval0                       -5.49332    0.35725 -15.377  < 2e-16 ***
## interval0.833333333333333       -1.58627    0.08934 -17.755  < 2e-16 ***
## interval1.25                    -0.71264    0.06996 -10.187  < 2e-16 ***
## interval1.75                    -0.19182    0.06824  -2.811  0.00494 **
## interval2.25                    -0.09650    0.07071  -1.365  0.17231
## interval3                       -0.90391    0.15313  -5.903 3.58e-09 ***
## interval4                       -1.77106    0.21238  -8.339  < 2e-16 ***
## interval5                       -2.10748    0.27635  -7.626 2.42e-14 ***
## interval6                       -3.23209    0.32672  -9.892  < 2e-16 ***
## interval9                       -3.32608    0.40059  -8.303  < 2e-16 ***
## interval12                      -5.96774    1.01191  -5.897 3.69e-09 ***
## age_groupyoung                   0.79675    0.16512   4.825 1.40e-06 ***
## factor(age_young_less_than_3)1 -0.46908    0.17696  -2.651  0.00803 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for poisson family taken to be 1)
```

Summary

- Parametric v data-driven v somewhere in between
- KM good if you have lots of data
- Parametric good if you have strong theory about what hazards should loook like
- I find PCH very useful as very flexible and easy to embed in Bayesian framework (more soon)
- Covariate-based proportional hazard models (PCH) very related to lots of ideas that we saw in GLM

At end of this slide deck, extra notes about Cox proportional hazards

- Not required, but very common so useful to know about

Extra: Cox proportional hazards

Cox proportional hazards

Here is an overview, as Cox PH is very common.

Consider a more general form of a proportional hazards model:

$$\lambda(t, \mathbf{x}) = \lambda_0(t, \alpha) \exp(\beta^T \mathbf{x})$$

- λ_0 is the baseline hazard, which depends on time but not the covariates
- $\exp(\beta^T \mathbf{x})$ depends on covariates but not time

If you are only interested in the effects of the covariates on survival, then you do not need to specify the form of the baseline hazard. Even without doing so you may estimate β.

Cox PH

Intuition: it all depends on the order.

First, consider two observations, T_1 and T_2 with hazard functions $\lambda_1(t)$ and $\lambda_2(t)$.

- The first failure happened at time t. What's the probability that this was subject 1?

$$Pr(T_1 = t | T_{(1)} = t) = \frac{\lambda_1(t)}{\lambda_1(t) + \lambda_2(t)} = Pr(T_1 < T_2)$$

We can rewrite this in terms of the proportional hazards, and the baseline hazard cancels out:

$$Pr(T_1 < T_2) = \frac{\exp(\mathbf{x}_1^T \beta)}{\exp(\mathbf{x}_1^T \beta) + \exp(\mathbf{x}_2^T \beta)}$$

Cox PH

In general,

- Order survival times $t_{(1)} < t_{(2)} < \cdots < t_{(n)}$.
- We may or may not censoring
- Define risk set $R(t)$ to be all the all the individuals still around at time t (i.e. not dead or censored)

Then the probability that subject j fails at time t given that one of the subjects from the risk set $R(t)$ failed at time t is

$$\frac{\exp(\mathbf{x}_j^T \beta)}{\sum_{k \in R(t)} \exp(\mathbf{x}_k^T \beta)}$$

Then an expression for the likelihood is

$$L(\beta) = \prod_j \frac{\exp(\mathbf{x}_j^T \beta)}{\sum_{k \in R(t_j)} \exp(\mathbf{x}_k^T \beta)}$$

where the product is taken over all observed survival times t_j

Cox PH

$$L(\beta) = \prod_j \frac{\exp(\mathbf{x}_j^T \beta)}{\sum_{k \in R(t_j)} \exp(\mathbf{x}_k^T \beta)}$$

Not really a likelihood in the true sense, called a **partial likelihood**

- not the full likelihood for α and β
- gives no information about probability of observing specific times t
- only uses relative ranking; actual times of events don't matter

Still, we can

- estimate β by solving $\frac{\partial \ell}{\partial \beta} = 0$ (most commonly, Newton-Raphson as before.)
- get SEs for βs from the inverse of the information matrix
- do the usual Wald, LR tests, etc

Cox PH in R

Note: very similar results to our PCH from before

```
fit_cox <- coxph(Surv(next.ivl, event)~age_group, data = f12)
summary(fit_cox)
```

```
## Call:
## coxph(formula = Surv(next.ivl, event) ~ age_group, data = f12)
##
##   n= 1840, number of events= 1657
##
##                    coef exp(coef) se(coef)     z Pr(>|z|)
## age_groupyoung 0.39294   1.48133  0.05966 6.586 4.51e-11 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##                exp(coef) exp(-coef) lower .95 upper .95
## age_groupyoung     1.481     0.6751     1.318     1.665
##
## Concordance= 0.529  (se = 0.006 )
## Likelihood ratio test= 46.44  on 1 df,   p=9e-12
## Wald test            = 43.38  on 1 df,   p=5e-11
## Score (logrank) test = 43.92  on 1 df,   p=3e-11
```

Testing the PH assumption

Schoenfeld residuals based on comparing the covariates x_i with their expected values + More details: Hosmer, chapter 6

Implement in R using:

```
cox.zph(fit_cox)
```

```
##              chisq df     p
## age_group  5.08  1 0.024
## GLOBAL     5.08  1 0.024
```

Evidence to reject PH assumption

Cox PH

Good:

- don't need to worry about specifying form of baseline hazard
- robust

Bad

- proportional hazards is pretty strong
- can fix by e.g. trying non-linear transforms of covariates, stratifying by covariates, but gets complex quick