Biostatistics III: Survival analysis for epidemiologists in R

Mark Clements
Department of Medical Epidemiology and Biostatistics
Karolinska Institutet
Stockholm, Sweden
http://www.biostat3.net/

4–13 November, 2024

https://doctoralcourses.application.ki.se/fubasextern/info?kurs=C8F2992, course code C8F2992

Topics for Day 4

- Generalised survival models (aka flexible parametric models)
- Non-collapsibility
- Nested case-control studies
- Standardised mortality (or incidence) ratios (SMR/SIR)
- Some possible biases in survival analysis
- Reporting on cohort studies

Splines I

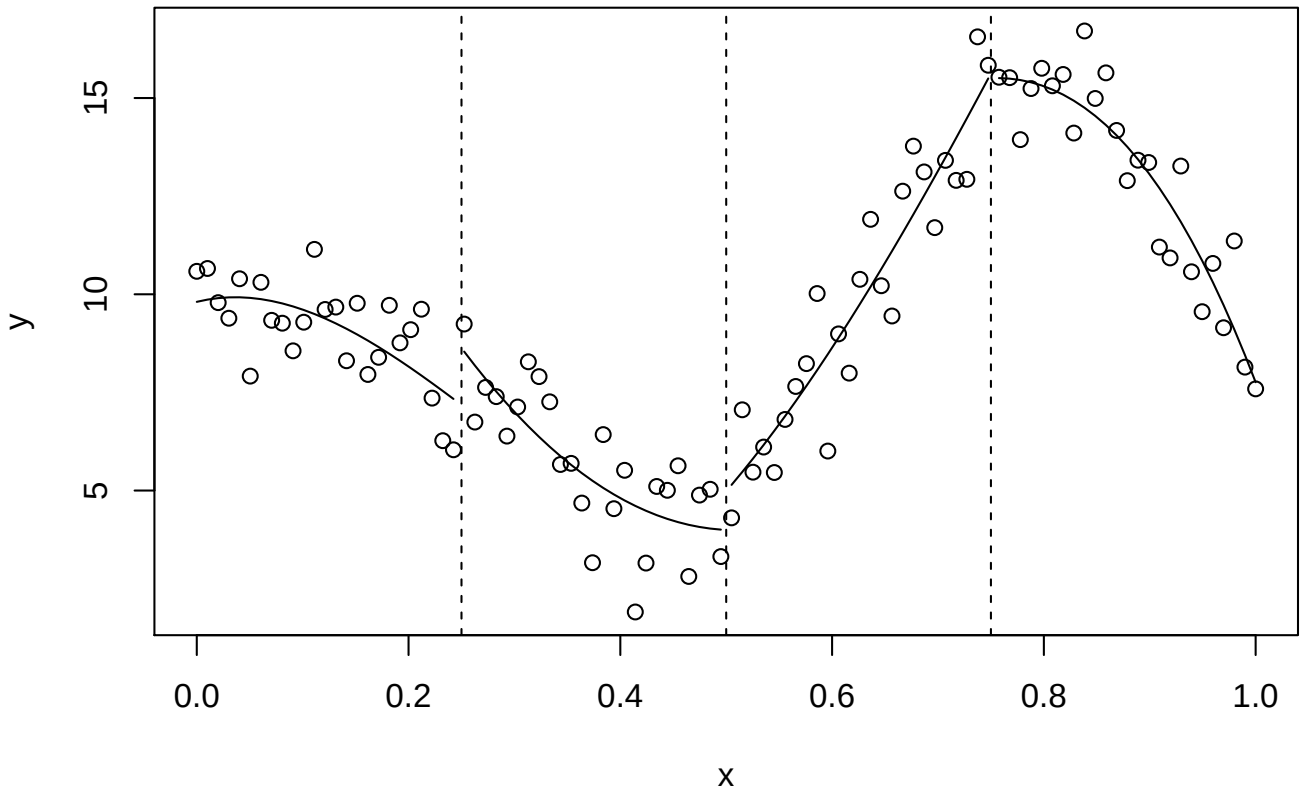
- In Poisson regression the baseline hazard is estimated as a step function.
- By fine splitting the steps can be made small and the baseline hazard approximately continuous. The drawback is that it requires the estimation of a lot of parameters.
- One alternative would be to use splines.
- Splines are a way of modeling continuous variables in a flexible way.

Splines II

- For cubic splines, we have piece-wise cubic functions for intervals defined by so-called knots
- For the knots, the user has to specify either
  1. The number of knots (e.g. df=3), with automatic knots placement (e.g. by quantiles of the event times);
  2. The number and placement of the knots; or
  3. Use some function that penalises the wiggliness of the function (so-called penalised regression)
- The user also has to specify the type of cubic splines
  1. B-splines, which are cubic before the first knot and after the last knot (`bs()` from the `splines` packages)
  2. Natural splines, which are linear before the first knot and after the last knot (`ns()`)
  3. Periodic splines
  4. Cure models (flat after the last knot)
  5. ...'
- For some interactive examples, see `https://pclambert.net/interactivegraphs/spline_eg/spline_eg`.

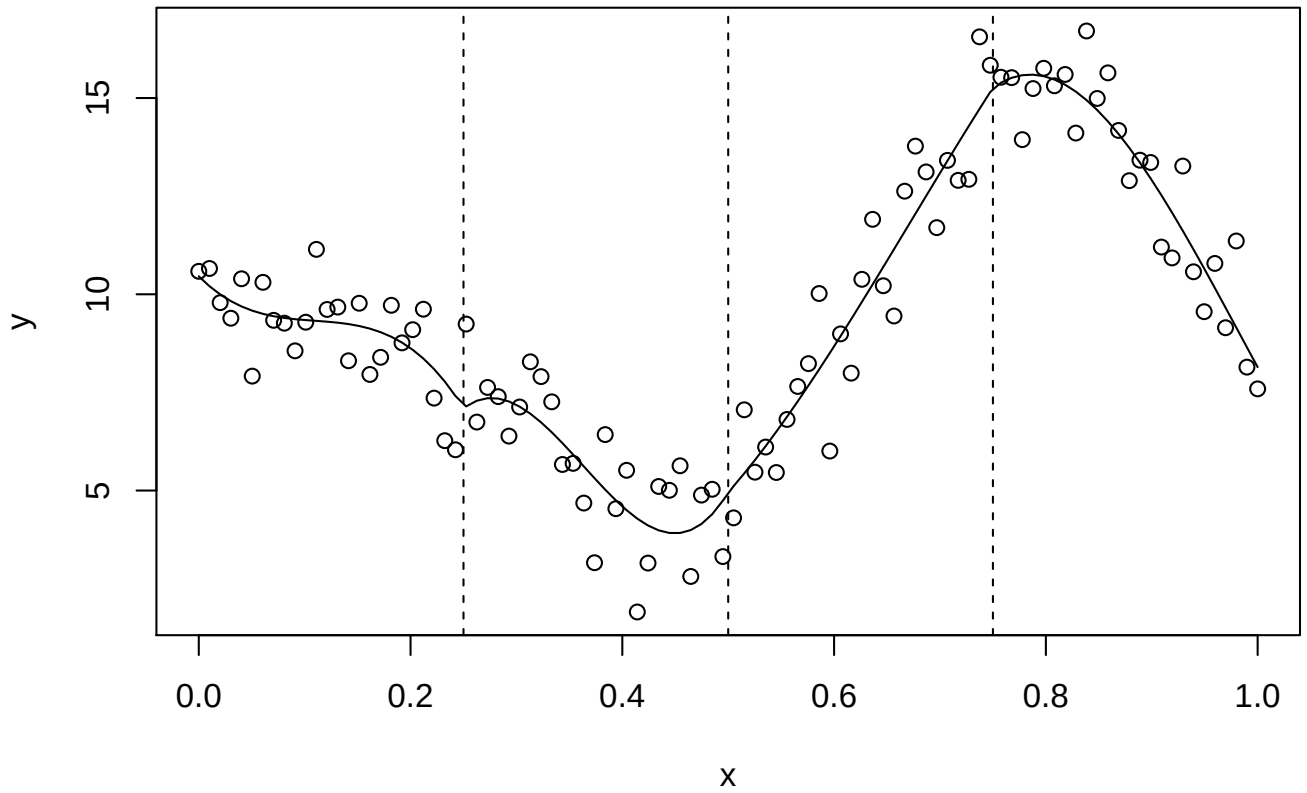
Splines III

Given data and some knots: initially assume no constraints



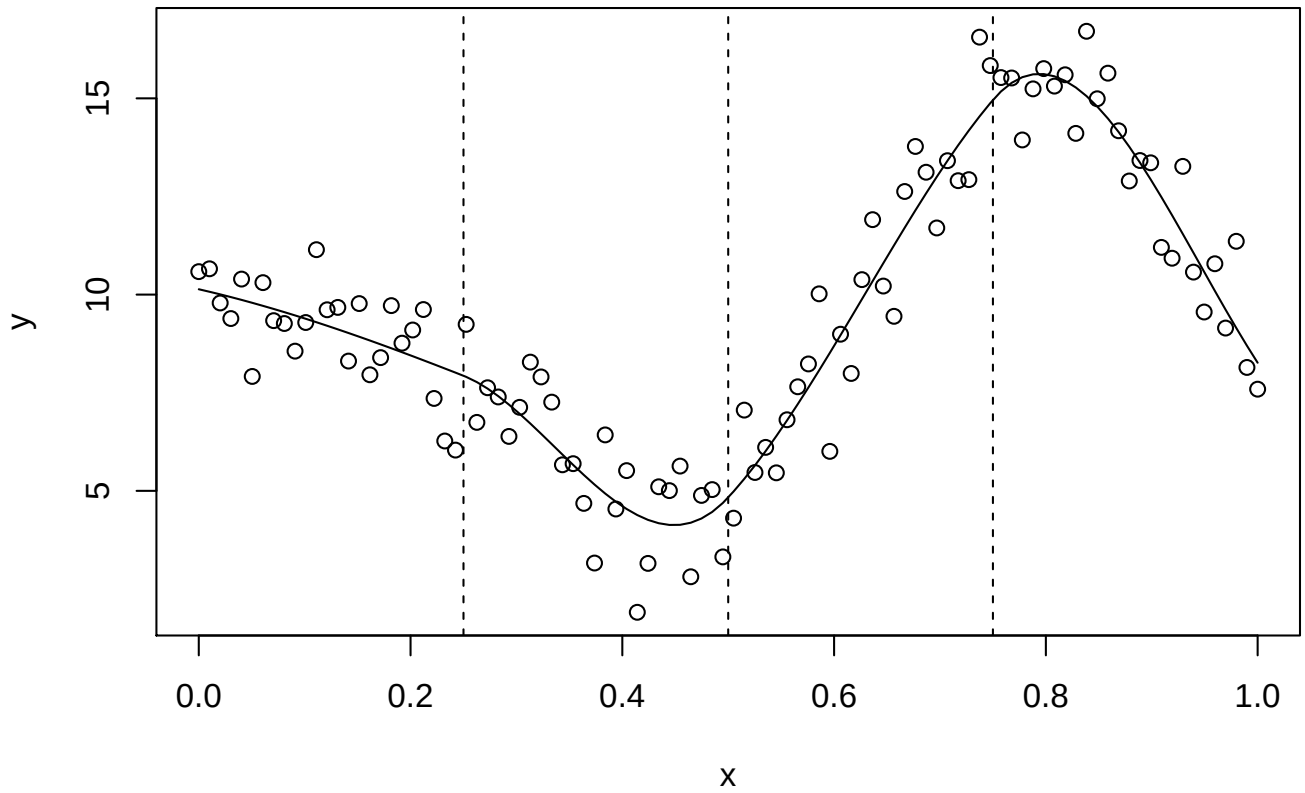
Splines IV

Now force the lines to join at knots



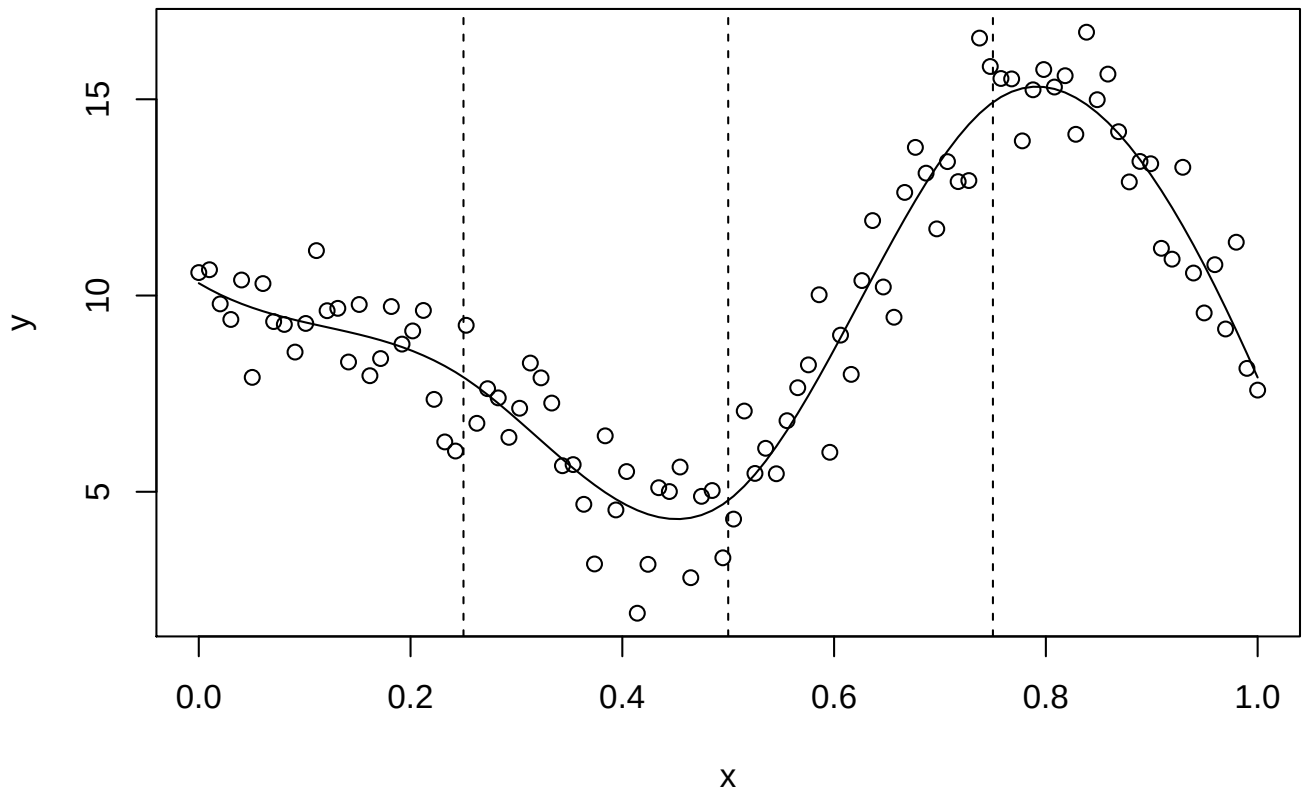
Splines V

Now assume continuous first derivatives



Splines VI

Now assume continuous first and second derivatives (B-splines)



Generalised survival models (aka flexible parametric survival models) I

- We can use splines to model for the baseline hazard
- One approach is to use Poisson regression and split finely for time. This works particularly well for multiple time scales.
- An alternative approach is to model for a transformation of survival – the so-called generalised survival models or flexible parametric survival models. The most common of these models is on the log cumulative hazard scale:

$$\log(\Lambda(t|\boldsymbol{x})) = s(\log t) + \beta^T \boldsymbol{x}$$

  where $s()$ is a smooth function (e.g. using natural splines) for log time.
- One advantage of the cumulative hazard scale is that we can easily calculate survival, such that $S(t|\boldsymbol{x}) = \exp(-\Lambda(t|\boldsymbol{x}))$
- We can calculate the hazard by differentiation, such that

$$\lambda(t|\boldsymbol{x}) = \Lambda'(t|\boldsymbol{x}) = \Lambda(t|\boldsymbol{x})s'(\log t)/t$$

Generalised survival models (aka flexible parametric survival models) II

- For the log cumulative hazard scale, we have proportional hazards:

$$\begin{aligned}\frac{\lambda(t|x = x_0 + 1)}{\lambda(t|x = x_0)} &= \frac{\Lambda(t|x = x_0 + 1)s'(\log t)/t}{\Lambda(t|x = x_0)s'(\log t)/t} \\ &= \frac{\exp(s(\log t) + \beta(x_0 + 1))}{\exp(s(\log t) + \beta(x_0))} \\ &= \exp(\beta)\end{aligned}$$

- This is a proportional hazards model, but non-proportional hazards (time-dependent effects) can be modeled by including interactions between covariates and splines for time. For example

$$\log(\Lambda(t|\boldsymbol{x})) = s(\log t) + \boldsymbol{\beta}^T\boldsymbol{x} + \sum_j s_j(\log t)x_j$$

where j is an index for the time-dependent effects.

Generalised survival models (aka flexible parametric survival models) III

- We have implemented both parametric and penalised generalised survival models in the `rstpm2` package on CRAN (see also the `flexsurv` package on CRAN and the `stpm2` command in Stata)
- The parametric models default to using natural splines for the smoothers by time. The investigator needs to specify the degrees of freedom (e.g. `df=3`)
- The penalised models do not need to specify the degrees of freedom. These models are closely related to generalised additive models, which can also be used with Poisson data.

Generalised survival models (aka flexible parametric survival models) IV

R code and output

```
> fit4 <- stpm2(Surv(surv_mm, status=="Dead: cancer") ~ sex + agegrp +
                      year8594,
                data=colon, subset=(stage=="Localised"), df=5)

                            poisson poisson.fine       coxph       stpm2
sexFemale                 -0.0929661  -0.08887086 -0.08939142 -0.08951965
agegrp45-59               -0.0502813  -0.05248568 -0.05198489 -0.05439771
agegrp60-74                0.2959041   0.29043188  0.29237391  0.28887930
agegrp75+                  0.8280071   0.80925760  0.81414446  0.81392895
year8594Diagnosed 85-94 -0.2789378  -0.28137264 -0.28254077 -0.27401607
```

Generalised survival models (aka flexible parametric survival models) V

- Hazard ratios are very similar to hazard ratios from a Cox model and Poisson model.
- Since the baseline hazard is modelled it is easy to include non-PH, interaction.
- The time-scale is included as a continuous variable, more plausible than step function.
- Easy to present results using graphs.
- The parametric approach enables predictions and extrapolations.

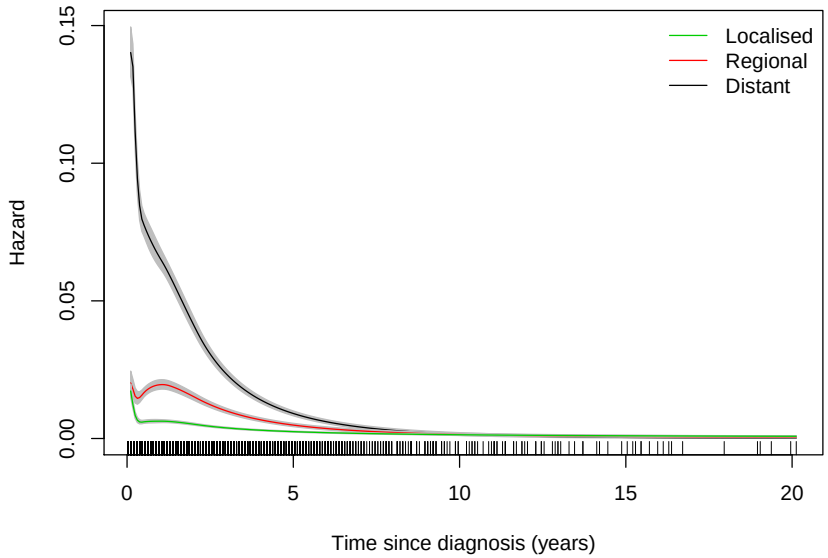
R code

```
known <- transform(colon,
                   distant=(stage=="Distant")+0,
                   regional=(stage=="Regional")+0,
                   stage=droplevels(stage,"Unknown"))
known <- subset(known, !is.na(stage))
```

Generalised survival models (aka flexible parametric survival models) VI

R code

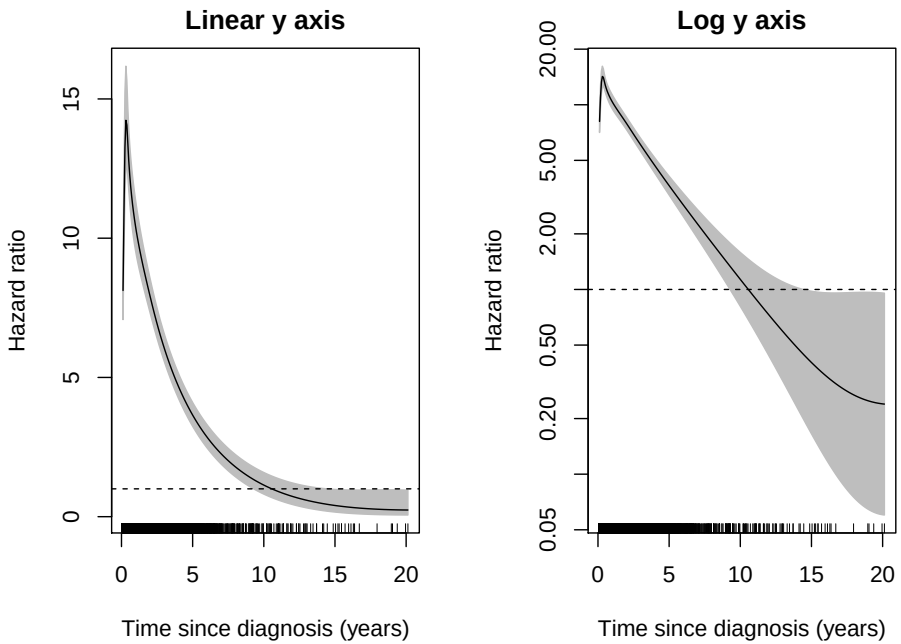
```
fit <- stpm2(Surv(surv_mm/12, status=="Dead: cancer") ~ sex + regional + distant,
             data=known, df=5, tvc=list(regional=3, distant=3))
plot(fit,type="hazard",newdata=data.frame(sex="Male",distant=1,regional=0),
     xlab="Time since diagnosis (years)")
lines(fit,type="hazard",newdata=data.frame(sex="Male",distant=0,regional=1),
      col=2, ci=TRUE)
lines(fit,type="hazard",newdata=data.frame(sex="Male",distant=0,regional=0),
      col=3, ci=TRUE)
legend("topright", legend=levels(known$stage), lty=1, col=3:1, bty="n")
```



Generalised survival models (aka flexible parametric survival models) VII

R code: Time-dependent hazard ratios comparing distant with localised colon cancer, males

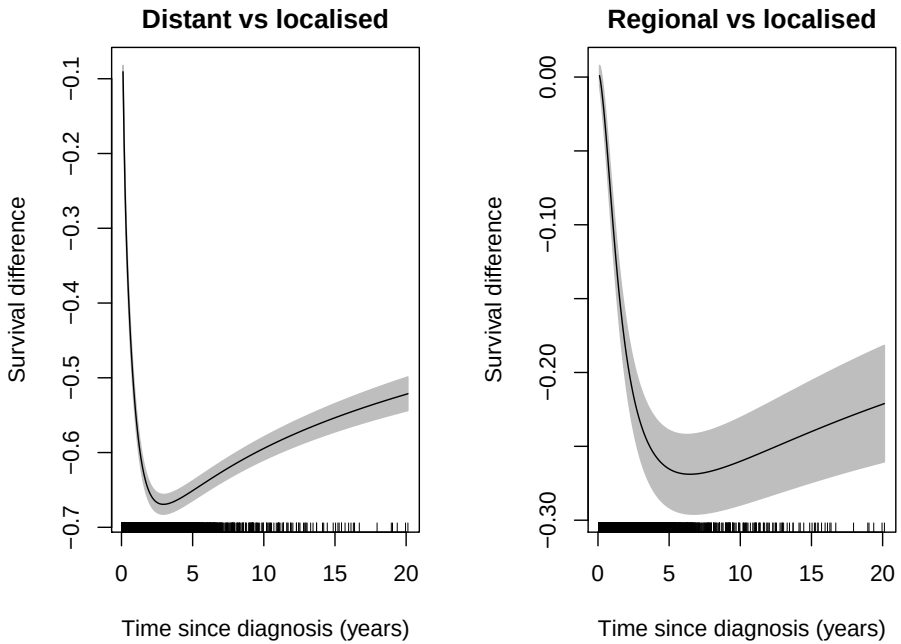
```
plot(fit,type="hr",newdata=data.frame(sex="Male",distant=0,regional=0),
     var="distant",
     xlab="Time since diagnosis (years)", main="Linear y axis")
```



Generalised survival models (aka flexible parametric survival models) VIII

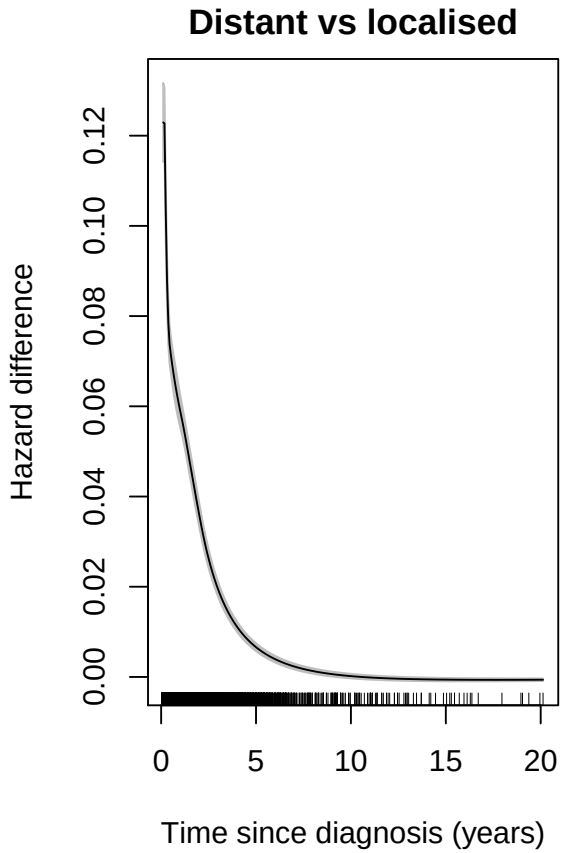
R code: Time-dependent survival differences comparing distant with localised colon cancer, males

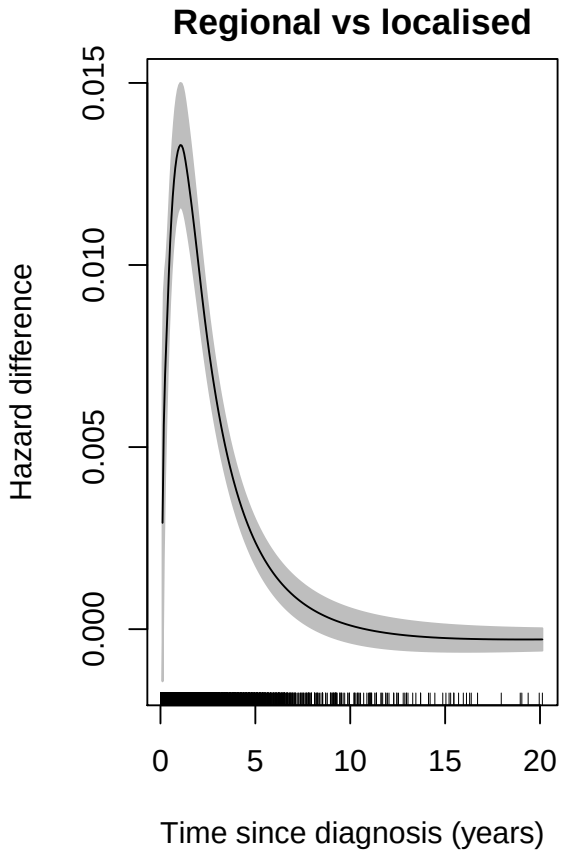
```
plot(fit,type="sdiff",newdata=data.frame(sex="Male",distant=0,regional=0),
     exposed=function(data) transform(data,distant=1),
     xlab="Time since diagnosis (years)", main="Distant vs localised")
```



Generalised survival models (aka flexible parametric survival models) IX

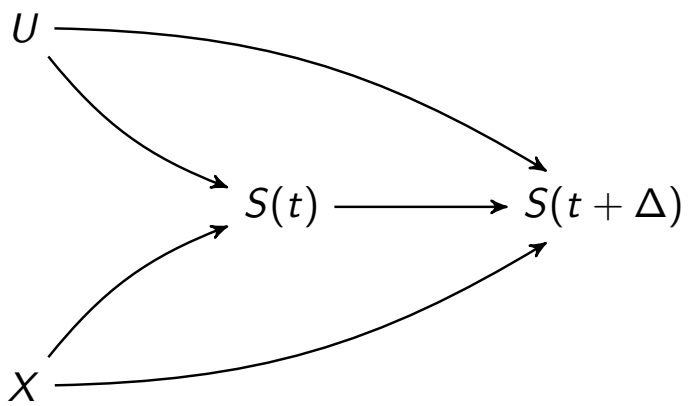
Time-dependent hazard differences comparing distant with localised colon cancer, males





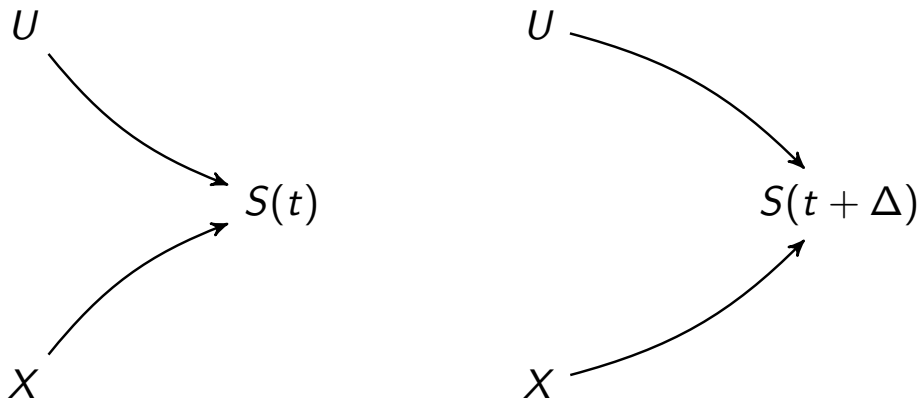
Proportional hazards and unmeasured covariates I

- What happens with proportional hazards models with unmeasured covariates?
- Assume that two covariates X and U are independent (e.g. for a randomised controlled trial, where X is the treatment effect and U are other covariates) and that both affect survival $S(t)$.
- Intuitively, the initial events for the Cox model will be randomised/independent, but later events are based on factors which affect survival – which will lead to a bias
- Consider the causal diagram for survival to time t and survival to time $t + \Delta$ (adapted from Aalen et al [1]):



Proportional hazards and unmeasured covariates II

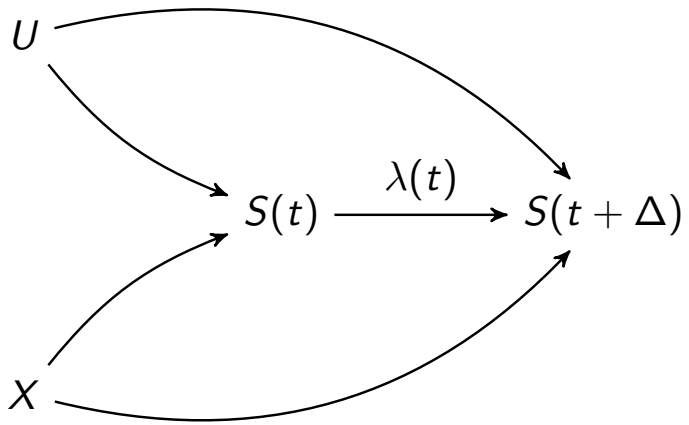
- If we consider how X affects survival at time t or time $t + \Delta$, then we do not open the path through U and we can get a causal estimator for X:



- This suggests that survival differences (or ratios) can have a causal interpretation

Proportional hazards and unmeasured covariates III

- As a reminder, the hazard between time t and time $t + \Delta$ is conditional on survival to time t. For hazard modelling, we have that:



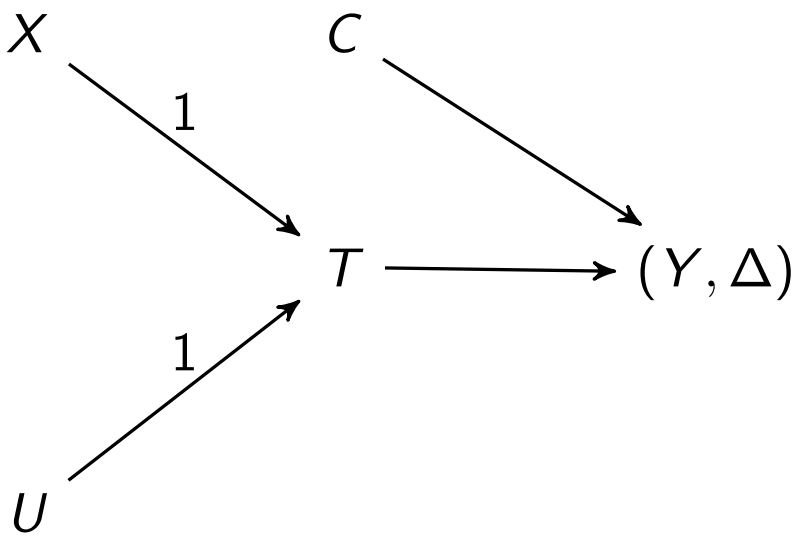
- When modelling for the hazard and adjusting for X, we implicitly adjust for $S(t)$, which acts as a collider.
- This opens the path from $X \rightarrow S(t) \leftarrow U \rightarrow S(t + \Delta)$.

Proportional hazards and unmeasured covariates

- What happens with proportional hazards models with unmeasured covariates?
- We consider the case when a covariate U is not associated with an exposure of interest X.
- A common approach in statistics and epidemiology is to start with a known truth and then see whether we can estimate a known target parameter from simulated data.
- For our simulation: simulate for a binary exposure X and a normally distributed covariate U; for simplicity, assume that the time to event T has hazards that are constant over time (that is, exponential) and the rate varies by X and U. Further assume that censoring C is uniform and independent of T.

Proportional hazards and unmeasured covariates

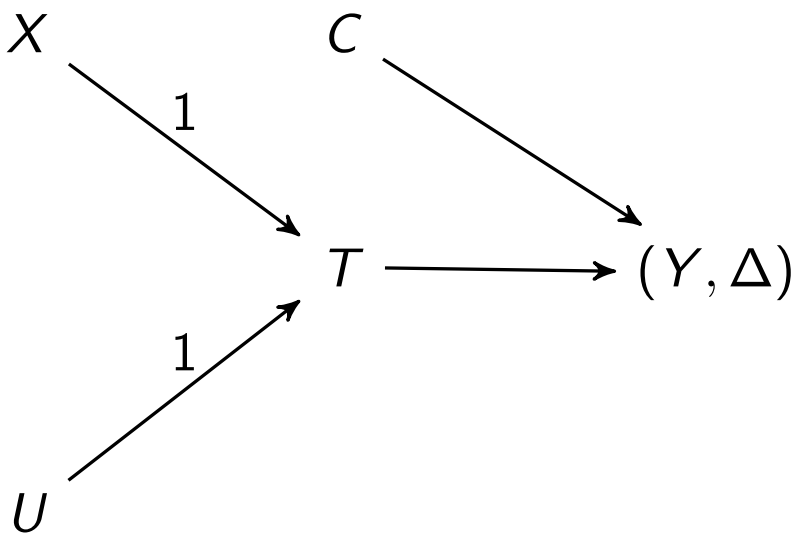
Given C and T, we can calculate the observed time Y and observed indicator Δ. The directed acyclic graph is then:



$$(Y, \Delta) = (\min(T, C),\ \ T < C)$$

Proportional hazards and unmeasured covariates

Given C and T, we can calculate the observed time Y and observed indicator Δ. The directed acyclic graph is then:



$$(Y, \Delta) = (\min(T, C),\ T < C)$$

where X is distributed as a Bernoulli variable with probability 0.5 (that is, a coin flip), U is normally distributed with mean 0 and standard deviation 3, T is exponentially distributed (that is, constant hazards with respect to time) with rate $\exp(-5 + X + U)$, and C is uniformly distributed between 0 and 10.

Simulation code

R code

```
set.seed(12345)
d <- local({
    n <- 1e4
    x <- rbinom(n, 1, 0.5)
    u <- rnorm(n, 0, 3)
    t <- rexp(n, exp(-5+x+u))
    c <- runif(n, 0, 10)
    y <- pmin(t, c)
    delta <- (t < c)
    data.frame(y,x,u,delta)
})
```

Simulation code

R code

```
set.seed(12345)
d <- local({
    n <- 1e4
    x <- rbinom(n, 1, 0.5)
    u <- rnorm(n, 0, 3)
    t <- rexp(n, exp(-5+x+u))
    c <- runif(n, 0, 10)
    y <- pmin(t, c)
    delta <- (t < c)
    data.frame(y,x,u,delta)
})
```

For discussion:

- How would you model with covariates x and u?
- How would you model with covariate x, when u is unmeasured?

Modelling for both x and u I

- There are many proportional hazards models that include constant hazards, including Poisson regression, Cox regression and flexible parametric survival models:

Modelling for both x and u II

R code and output

```
> fit1 <- glm(delta~x+u+offset(log(y)),data=d,
              family=poisson)
> fit2 <- coxph(Surv(y,delta)~x+u,data=d)
> fit3 <- stpm2(Surv(y,delta)~x+u,data=d,df=4)

> rbind(Poisson=coef(summary(fit1))["x",c("Estimate",
                                          "Std. Error")],
        Cox=coef(summary(fit2))["x",c("coef","se(coef)")],
        Stpm2=coef(summary(fit3))["x",c("Estimate",
                                        "Std. Error")])

          Estimate Std. Error
Poisson 0.9602659 0.04372345
Cox     0.9622728 0.04490412
Stpm2   0.9607498 0.04487831
```

Modelling for both x and u III

In summary:

- All three models estimate the target log hazard ratio for X (≈ 1).
- The standard error for Poisson is similar to and slightly smaller than those for Cox and stpm2

Modelling for only x, without u I

Modelling for only x, without u II

R code and output: Using the same model classes and excluding u from the models

```
> fit1 <- glm(delta~x+offset(log(y)),data=d,
              family=poisson)
> fit2 <- coxph(Surv(y,delta)~x,data=d)
> fit3 <- stpm2(Surv(y,delta)~x,data=d,df=4)

> rbind(Poisson=coef(summary(fit1))["x",c("Estimate",
                                          "Std. Error")],
        Cox=coef(summary(fit2))["x",c("coef","se(coef)")],
        Stpm2=coef(summary(fit3))["x",c("Estimate",
                                        "Std. Error")])

          Estimate Std. Error
Poisson 0.5128869 0.04341441
Cox     0.4851147 0.04342341
Stpm2   0.4852363 0.04342328
```

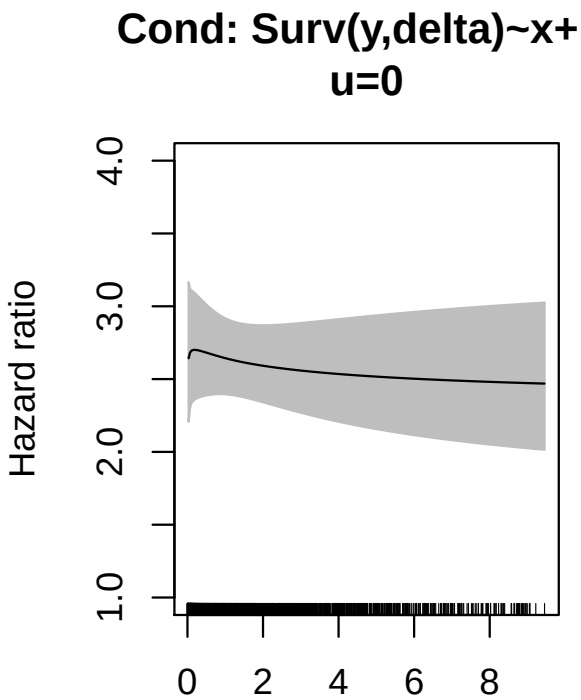
Modelling for only x, without u III

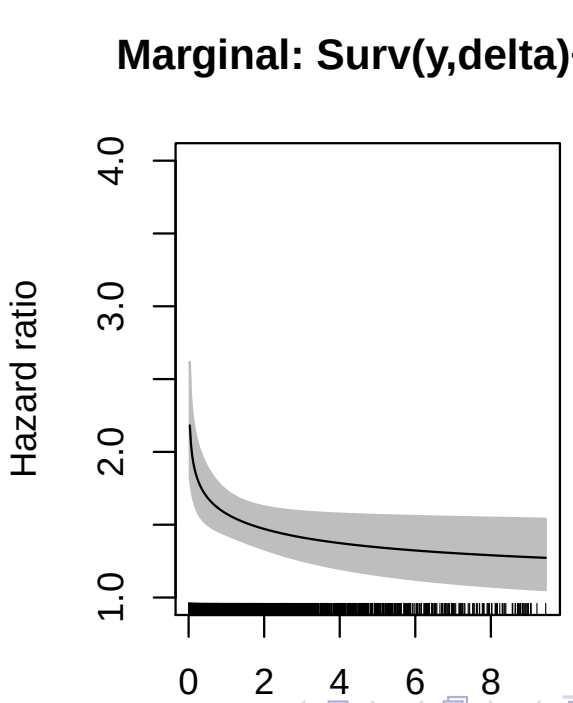
- The marginal hazard ratio can be quite different to the conditional hazard ratio from the previous models
- As an exercise (see Exercise 14), we can investigate whether the hazard ratios are time-varying for the stpm2 models with both x and u or with only x (see next)

Marginal and conditional time-varying hazard ratio for x

R code

```
> fit <- stpm2(Surv(y,delta)~x+u,data=d, df=4, tvc=list(x=2))
> plot(fit, type="hr", newdata=data.frame(x=0,u=0), var="x",
      ylim=c(1,4), main="Cond: Surv(y,delta)~x+u,\nu=0")
> fit <- stpm2(Surv(y,delta)~x,data=d, df=4, tvc=list(x=2))
> plot(fit, type="hr", newdata=data.frame(x=0), var="x",
      ylim=c(1,4), main="Marginal: Surv(y,delta)~x")
```





Why is the marginal estimate different to the conditional estimate?

- The hazard ratio for X for the model with both x and u is conditional on the values for u. For our example, the hazard ratio is constant with respect to time.
- The hazard ratio for X for the model with only x is marginal (or an average) over the values for u. For our example, the hazard ratio starts close to exp(1) and is then attenuated
- If we fit a marginal model with a constant hazard ratio, we estimate an average of the time-varying hazard ratio
- Individuals with higher values of u and x are more likely to have an event earlier, so that there is selection. At later times, the distributions of U and X for the survivors are expected to have lower values.
- For normally distributed unmeasured covariates U, theory says that the marginal hazard ratio for X will be attenuated
- The unmeasured covariate U (or unmeasured heterogeneity) is also called a random effect, and $\exp(U)$ is called a statistical frailty

Non-collapsibility of the hazard ratio

- We would prefer to have a model which is collapsible, where the marginal estimate without the unmeasured covariates is the same parameter as the conditional estimate
  - A related example is for odds ratios, which are also not collapsible
- To emphasise, we assumed that U was not a confounder for X. Consequently, this issue is similar for both randomised controlled trials and observational data.
- We can reduce the attenuation by modelling for covariates that are not confounders — but we shouldn't need to do this!
- For proportional hazards there are two conditions for HR $\neq 1$ where the non-collapsibility is negligible. . .

Simulation code: rare events

R code

```
set.seed(12345)
d <- local({
    n <- 1e4*10 # CHANGED N
    x <- rbinom(n, 1, 0.5)
    u <- rnorm(n, 0, 3)
    t <- rexp(n, exp(-5+x+u))
    c <- runif(n, 0, 10/1000) # CHANGED FROM 10 TO 0.01
    y <- pmin(t, c)
    delta <- (t < c)
    data.frame(y,x,u,delta)
})
```

Modelling for only x, without u (rare events) I

- Using the same model classes and excluding u from the models:

R code and output

```
> fit1 <- glm(delta~x+offset(log(y)),data=d,
              family=poisson)
> fit2 <- coxph(Surv(y,delta)~x,data=d)
> fit3 <- stpm2(Surv(y,delta)~x,data=d,df=4)
> rbind(Poisson=coef(summary(fit1))["x",c("Estimate",
                                          "Std. Error")],
        Cox=coef(summary(fit2))["x",c("coef","se(coef)")],
        Stpm2=coef(summary(fit3))["x",c("Estimate",
                                        "Std. Error")])

         Estimate Std. Error
Poisson 1.081150  0.1212037
Cox     1.081513  0.1212136
Stpm2   1.081399  0.1212136
```

Simulation code: smaller frailty

R code

```
set.seed(12345)
d <- local( {
    n <- 1e4
    x <- rbinom(n, 1, 0.5)
    u <- rnorm(n, 0, 1) # CHANGED SD FROM 3 TO 1
    t <- rexp(n, exp(-5+x+u))
    c <- runif(n, 0, 10)
    y <- pmin(t, c)
    delta <- (t < c)
    data.frame(y,x,u,delta)
})
```

Modelling for only x, without u (smaller frailty) I

- Using the same model classes and excluding u from the models:

R code and output

```
> fit1 <- glm(delta~x+offset(log(y)),data=d,
              family=poisson)
> fit2 <- coxph(Surv(y,delta)~x,data=d)
> fit3 <- stpm2(Surv(y,delta)~x,data=d,df=4)
> rbind(Poisson=coef(summary(fit1))["x",c("Estimate",
                                          "Std. Error")],
        Cox=coef(summary(fit2))["x",c("coef","se(coef)")],
        Stpm2=coef(summary(fit3))["x",c("Estimate",
                                        "Std. Error")])

          Estimate Std. Error
Poisson 0.9498227 0.07583196
Cox     0.9444216 0.07584759
Stpm2   0.9441747 0.07584672
```

Summary (so far) I

- Non-collapsibility is less of an issue with (i) rare events and (ii) a small frailty
- Unfortunately, for single, unclustered events we cannot assess the size of the frailty — modelling indirectly helps, but we still cannot characterise the unmodelled frailty

Approaches to address non-collapsibility I

- First, we can use regression standardisation to calculate marginal effects from the conditional model. This includes standardised hazard ratios (which look similar to the marginal hazard ratios) and standardised survival differences
- Standardised survival $\bar{S}(t)$ at time t is calculated by

$$\bar{S}(t) = \sum_i w_i S(t|\boldsymbol{x}_i)$$

for covariate patterns indexed by i with weights w_i and covariates $\boldsymbol{x}_i$. If we standardise by the observed data, we have

$$\bar{S}(t) = \sum_{i=1}^{n} \frac{1}{n} S(t|\boldsymbol{x}_i)$$

where i is an index over the observations.

Approaches to address non-collapsibility II

- We can estimate standardised survival under counterfactual exposures. For example, we could define

$$\bar{S}_k(t) = \sum_{i=1}^{n} \frac{1}{n} S(t|u_i, \text{do}(x = k))$$

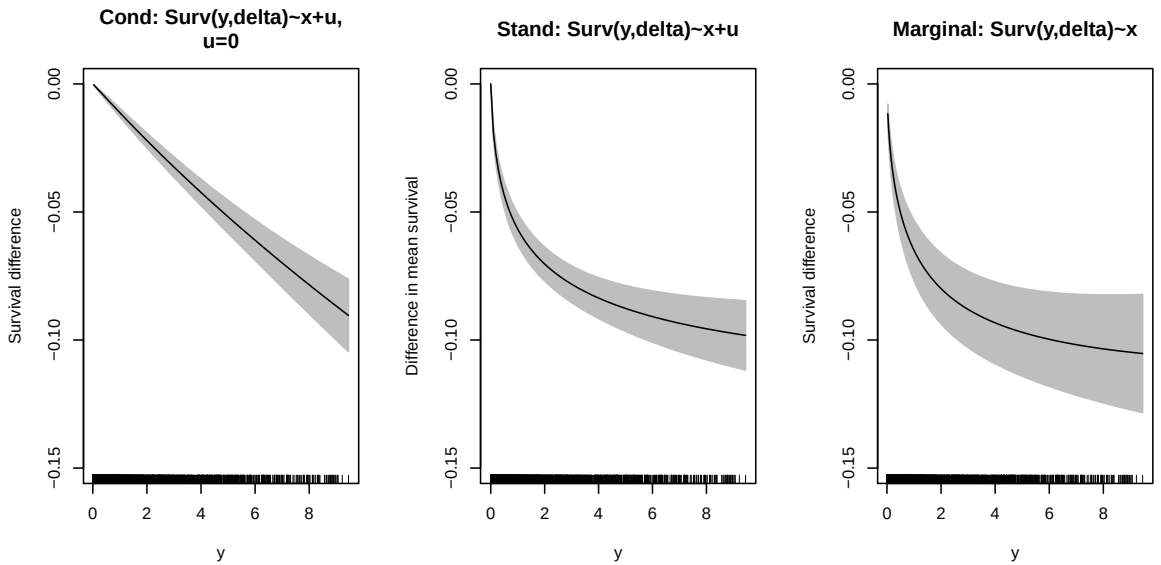
which sets all observations to having $x = k$, and then the standardised survival difference would be

$$\bar{S}_1(t) - \bar{S}_0(t) = \sum_{i=1}^{n} \frac{1}{n} \left(S(t|u_i, \text{do}(x = 1)) - S(t|u_i, \text{do}(x = 0)) \right)$$

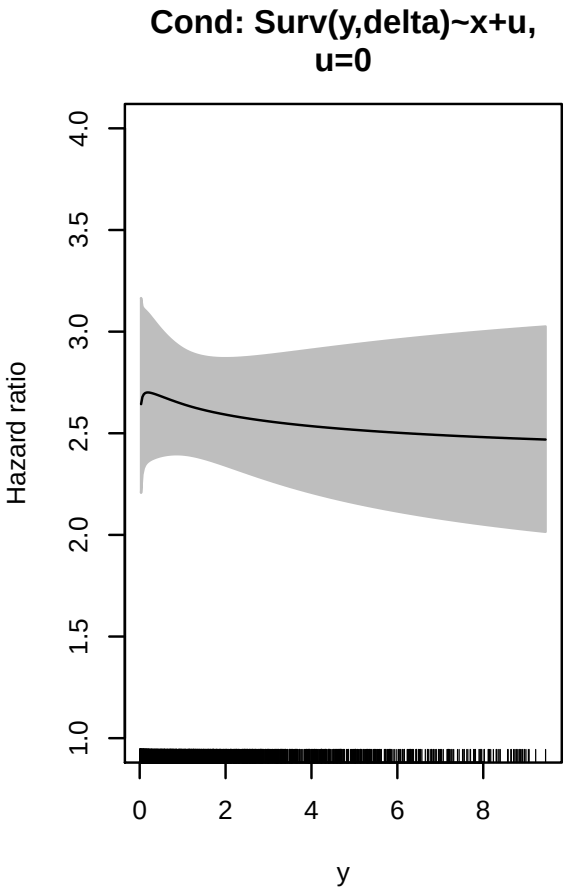
Survival differences

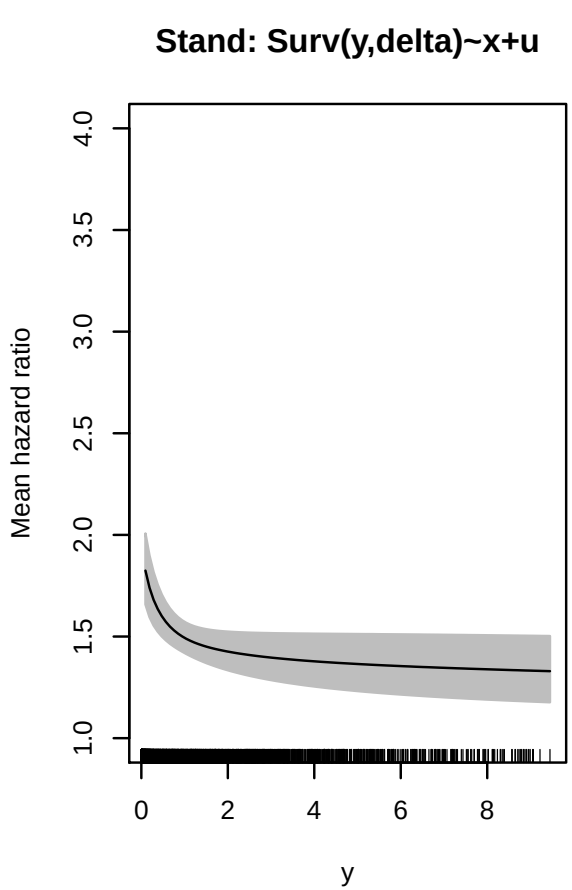
R code

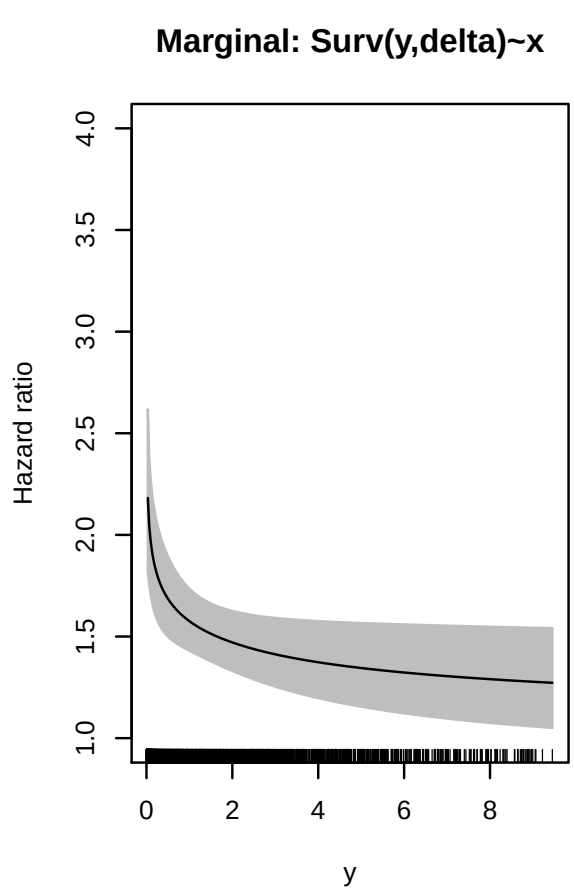
```
fit <- stpm2(Surv(y,delta)~x+u,data=d,df=4, tvc=list(x=2))
plot(fit, type="sdiff", newdata=data.frame(x=0,u=0), var="x",
     ylim=c(-0.15,0), main="Cond: Surv(y,delta)~x+u,\nu=0")
plot(fit, type="meansurvdiff", newdata=transform(d,x=0), var="x",
     ylim=c(-0.15,0), main="Stand: Surv(y,delta)~x+u")
fit <- stpm2(Surv(y,delta)~x,data=d,df=4, tvc=list(x=2))
plot(fit, type="sdiff", newdata=data.frame(x=0), var="x",
     ylim=c(-0.15,0), main="Marginal: Surv(y,delta)~x")
```



Hazard ratios







Other approaches to address this non-collapsibility I

- Second, we could use survival models that *are* collapsible
- A popular choice has been Aalen's additive hazards model, where (assuming a continuous hazard)

$$\lambda(t|\boldsymbol{x}) = \lambda_0(t) + \sum_j x_j \beta_j(t)$$

- Although this is a hazards model, the linearity means that the marginal and conditional models are similar for covariates that are not confounders.
- The different exposures are assumed to act like competing events, similar to Rothman's causal pies, rather than proportionally. This additivity does not seem to hold for some event types (e.g. cancer survival by age and sex).
- These models are implemented in the timereg package on CRAN. The models are reported using non-parametric, time-varying cumulative effects, such that

$$\Lambda(t|\boldsymbol{x}) = \int_0^t \lambda(u|\boldsymbol{x})du = \int_0^t \lambda_0(u)du + \sum_j x_j \int_0^t \beta_j(u)du$$

- This model does not hold for our simulated data.

Other approaches to address this non-collapsibility II

- An alternative approach, which shows considerable promise, is to use accelerated failure time models, where $T = T_0 \exp(-\beta x)$, which is proportionality on the time scale.
- We have recently developed smooth accelerated failure time models which allow some flexibility in the baseline survival, such that

$$S(t|x) = S_0(t \exp(-\beta x))$$

for some flexible $S_0(t)$ (e.g. splines for `rstpm2::aft`).

Other approaches to address this non-collapsibility III

- Our simulated data, which are exponential, can also be fitted as an accelerated failure time model:

R code

```
fit.xu <- aft(Surv(y,delta)~x+u,data=d,df=4)
fit.x <- aft(Surv(y,delta)~x,data=d,df=4)
rbind("Surv(y,delta)~x+u"=coef(summary(fit.xu))["x",c("Estimate",
                                                      "Std. Error")],
      "Surv(y,delta)~x"=coef(summary(fit.x))["x",c("Estimate",
                                                   "Std. Error")])
```

```
                     Estimate Std. Error
Surv(y,delta)~x+u -0.962349 0.04409720
Surv(y,delta)~x   -1.036672 0.09105875
```

- We see that the AFT model is collapsible.

Likelihood calculations for the Cox model I

- Estimation is based on the concept of *risk sets*. Understanding this is central to understanding risk set sampling (e.g. nested case-control and case-cohort studies) presented on slide 50.
- The risk set at each failure time is the collection of subjects who were at risk of failing at that time.
- In theory, only one individual can fail at each failure time and we can calculate the conditional probability of failure for the subject who actually failed.
- The partial likelihood function is the product of these conditional probabilities.
- Imagine 5 individuals at risk at time t of which one fails.
- These individuals have hazards $\lambda_1, \lambda_2, \ldots, \lambda_5$ which may be different since the individuals have different covariate values.
- Conditional on one of the five failing, the probability it is number 2 is

$$\frac{\lambda_2}{\lambda_1 + \lambda_2 + \lambda_3 + \lambda_4 + \lambda_5}$$

Likelihood calculations for the Cox model II

- Since $\lambda(t) = \lambda_0(t)\exp(x\beta)$ we can write this as

$$\frac{\lambda_0(t)\exp(x_2\beta)}{\lambda_0(t)\exp(x_1\beta) + \lambda_0(t)\exp(x_2\beta) + \ldots + \lambda_0(t)\exp(x_5\beta)}$$

- The baseline hazard, $\lambda_0(t)$, cancels and we have

$$\frac{\exp(x_2\beta)}{\sum_{i\in R}\exp(x_i\beta)}$$

  where R represents the risk set.
- The likelihood function is the product of these conditional probabilities.
- If we have J distinct failure times then

$$L(\beta) = \prod_{j=1}^{J}\left(\frac{\exp(x_j\beta)}{\sum_{i\in R_j}\exp(x_i\beta)}\right)$$

Likelihood calculations for the Cox model III

- The likelihood function can be further generalised for K (potentially time-dependent) covariates $x_{i1}(t), \ldots, x_{iK}(t)$

$$L(\beta) = \prod_{j=1}^{J} \left(\frac{\exp\left(\sum_{k=1}^{K} \beta_k x_{jk}(t_j)\right)}{\sum_{i \in R_j} \exp\left(\sum_{k=1}^{K} \beta_k x_{ik}(t_j)\right)} \right) \tag{1}$$

- Note that these calculations do not depend on the underlying failure times; only the ordering of failure times is important.
- Although this is not a likelihood in the strict sense, it is a partial likelihood, it can for all intents and purposes be treated as a likelihood.
- In practice we often observe multiple failures at the same time (ties) and need to use an approximation to equation 1.
- Conceptually similar to a matched (on time) case-control study. Cox partial likelihood is similar to the likelihood for conditional logistic regression (used for analysing matched case-control studies).

Sampling from the risk set: the nested case-control design I

- When fitting the Cox model we essentially compare, at every failure time, the characteristics of the individual who failed to the characteristics of all individuals who did not fail (equation 1).
- We could think of this as a case-control study matched on time; at each failure time we have one case and several hundred (or more) controls.
- We could instead select, for example, 5 controls per case with little loss of efficiency.
- Our controls are selected from the risk set; a single individual may be a control at multiple time points and a control may later become a case.
- This is a nested case-control design; a case-control study nested within a cohort.
- This design has become popular because it allows for statistically efficient analysis of data from a cohort with substantial savings in cost and time.
- We may wish, for example, to extract information from medical records for the patients diagnosed with colon carcinoma in order to study additional explanatory variables.

Sampling from the risk set: the nested case-control design II

- This would be an ideal setting for a nested case-control design; we extract information for all individuals who died but only a sample of those who did not.
- Another ideal application is where we establish a population-based cohort and take blood samples with the aim of studying the association between genotype and disease risk.
- We store the blood samples and only after following up the cohort do we analyse the samples for the cases (individuals who developed the disease) along with a sample of controls.
- Generating a nested case-control study is very easy in R. First, however, we'll repeat the full cohort analysis of the localised colon carcinoma data.

Sampling from the risk set: the nested case-control design III

R code and output

```
> localised <- subset(colon, stage == "Localised")
> localised <- transform(localised,
                         event=(status=="Dead: cancer"))
> summary(coxph(Surv(surv_mm, event) ~ sex+agegrp+year8594,
                data=localised))

  n= 6274, number of events= 1734

                            coef exp(coef) se(coef)      z Pr(>|z|)
sexFemale                -0.08939   0.91449  0.04937 -1.811   0.0702 .
agegrp45-59              -0.05198   0.94934  0.13845 -0.375   0.7073
agegrp60-74               0.29237   1.33960  0.12573  2.325   0.0201 *
agegrp75+                 0.81414   2.25724  0.12607  6.458 1.06e-10 ***
year8594Diagnosed 85-94 -0.28254   0.75387  0.04937 -5.723 1.05e-08 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Sampling from the risk set: the nested case-control design IV

- We will generate a nested case-control study with three controls per case matching on age group.

R code and output

```
> library(Epi)
> set.seed(12345)
> cc <- Epi::ccwc(exit=surv_mm, fail=event,
                  data=localised, include=list(sex,year8594),
                  controls=3, match=agegrp, silent=TRUE)
> tail(cc)

 Warning message:
In Epi::ccwc(exit = surv_mm, fail = event, data = localised, include = list(sex,  :
  there were tied failure times
     Set  Map Time Fail agegrp    sex        year8594
6931 353 2406 73.5    0   0-44 Female Diagnosed 75-84
6932 353 3600 73.5    0   0-44 Female Diagnosed 85-94
6933 354 3792 49.5    1   0-44 Female Diagnosed 85-94
6934 354 1335 49.5    0   0-44   Male Diagnosed 75-84
6935 354 3527 49.5    0   0-44   Male Diagnosed 85-94
6936 354 2948 49.5    0   0-44   Male Diagnosed 85-94
```

Sampling from the risk set: the nested case-control design V

- The resulting nested case-control study is analysed using conditional logistic regression (theoretically very similar to Cox regression).

R code and output

```
> clogit(Fail ~ sex+year8594+strata(Set), data=cc)

Call:
clogit(Fail ~ sex + year8594 + strata(Set), data = cc)

                          coef exp(coef) se(coef)     z       p
sexFemale                -0.1085    0.8971   0.0571 -1.90   0.057
year8594Diagnosed 85-94 -0.2914    0.7472   0.0571 -5.11 3.3e-07

Likelihood ratio test=29.1  on 2 df, p=4.85e-07
n= 6936, number of events= 1734
```

- Estimates are similar to the full cohort but standard errors are slightly higher.

Standardised mortality/incidence ratios I

- Sometimes you only have an exposed population, and no unexposed population to compare with.
- We can then not follow-up both the exposed and the unexposed population and compare the two estimated rates.
- We instead only estimate the rate (or number of events) in the exposed population and compare this to the expected rate (expected number of events) for the standard population.
- For example, we might study disease incidence or mortality among individuals with a certain occupation (farmers, painters, airline cabin crew) or cancer incidence in a cohort exposed to ionising radiation.
- The standardized mortality ratio (SMR) is the ratio of the observed number of deaths in the study population to the number that would be expected if the study population experienced the same mortality as the standard population.
- It is an indirectly standardized rate.
- When studying disease incidence the corresponding quantity is called a standardized incidence ratio (SIR).

Standardised mortality/incidence ratios II

- Example, estimating relative risk of cancer among organ transplant recipients compared to the general population [4].

Estimation in R I

- To estimate an SIR/SMR, we typically assume the expected rates depend on age, sex, and calendar period.
- We use the `survSplit` function to split follow-up time in the same way as the file of background rates is classified. That is, if the population rates are for 5-year age groups we split using the same categories.
- Multiply the expected rates by the person-times, Y, to get the expected number of events in each category.
- Collapse to get the total number of observed (O) and expected (E) events and calculate the standardised incidence ratio SIR=O/E.
- Modelling is done in the same way as for rates except we have $\log(E)$ as the offset instead of $\log(Y)$.

Estimation in R II

Example of a binary exposure x with collapsed data

R code and output

```
  ## example
> sir <- data.frame(x=c(0,1), O=c(21,32), E=c(10,10))
> eform(glm(O ~ x + offset(log(E)), data=sir, family=poisson))
            exp(beta)     2.5 %   97.5 %
(Intercept)   2.10000 1.3692158 3.220822
x             1.52381 0.8787844 2.642281
```

Estimation in R III

As a second example, consider a cohort study of men exposed to household asbestos from the Australian Capital Territory. There were 7 male cases, with 2.75 expected (Korda et al 2017)

R code and output

```
> poisson.test(x=7, T=2.75)

 Exact Poisson test

data:  7 time base: 2.75
number of events = 7, time base = 2.75, p-value = 0.02243
alternative hypothesis: true event rate is not equal to 1
95 percent confidence interval:
 1.023405 5.244609
sample estimates:
event rate
  2.545455
```

Selection bias in observational studies I

- The aim of a study is usually to derive from an available subset of patients, statements about their patterns of survival which will be generalisable to a wider body of patients.
- Selection bias can occur when patients treated at a given clinic are not representative of a general class of patients. For example, if seriously ill patients are transferred to a specialist clinic then neither patients treated at the 'general' clinic or the specialist clinic will be representative of 'all patients'.
- Selection bias also occurs when treatment is assigned based on characteristics of the patients, thereby precluding comparisons between treatment groups.
- Patients treated aggressively are generally healthier than patients treated conservatively. For example:
  (i) Radical prostatectomy vs 'watchful waiting' (expectant therapy) for men diagnosed with localised prostate cancer.
  (ii) Bone marrow transplant and high dose chemotherapy vs conventional therapies for women diagnosed with advanced breast cancer.
- High-dose chemotherapy accompanied by transplant involves harvesting bone marrow or stem cells from the patient prior to chemotherapy.

Selection bias in observational studies II

- The patient then receives high-dose chemotherapy, which adversely affects bone marrow. After chemotherapy, the stem cells and bone marrow are replaced with the hope that the drugs have killed the cancer cells and the bone marrow will regenerate before the patient dies of infection.
- The procedure was started in 1979 and by the late 1980s the results looked very promising. Patients with advanced breast cancer who were given transplants had remission rates of 50 to 60 percent compared with the 10 to 15 percent remission rates achieved by conventional means.
- Subsequent examination of the data showed that the women receiving the transplant treatment were carefully selected to be younger than 60 and in general good health.
- Recently completed randomised clinical trials found no difference in survival for women who were randomly assigned to have transplants and those who were assigned to conventional therapy.
- In general, comparison of survival according to treatment should be avoided in observational studies.

Selection bias in observational studies III

- It is possible to adjust for factors which make a patient subgroup atypical (e.g. disease characteristics, presence of comorbid conditions, age, etc.) but there is no substitute for a randomised experimental trial for evaluating different treatments.
- For causal inference using observational data, we may be interested in emulating a target randomised trial (see later)

Another problem which arises when comparing treatments

- A common question is whether a combination treatment (e.g. surgery followed by radiation therapy) is preferable to the single treatment (e.g. surgery alone).
- Survival time is usually measured from date of diagnosis, date of first hospital admission, or date of first treatment.
- In order to receive the combination treatment, one must survive a sufficient period after surgery in order to receive the radiation therapy.
- Those who die during, or immediately after, surgery are included in the 'surgery only' group.
- A naive analysis would show that the group receiving combination therapy experience superior survival.

Risk prediction

- Risk prediction relates to the calculation of the failure function.

$$\mathrm{Risk}(t; \boldsymbol{x}) = 1 - S(t; \boldsymbol{x}) = 1 - \exp\left(-\int_0^t \lambda(u; \boldsymbol{x}) \mathrm{d}u\right)$$

- If the hazard λ is constant, then $\mathrm{Risk}(t) = 1 - \exp(-\lambda t)$.
- For Cox regression, the baseline survival $S_0(t)$ is estimated using the Breslow estimator. This can be estimated using the `predict` statement. For a linear predictor $\beta^T \boldsymbol{x}$,

$$\mathrm{Risk}(t; \boldsymbol{x}) = 1 - S_0(t)^{\exp(\beta^T \boldsymbol{x})}$$

- Standard errors and confidence intervals for the risk estimates from Cox regression can be calculated by reparameterising the model such that $\boldsymbol{x} = 0$ represents the covariates of interest.

Guidelines for publishing/presenting survival studies

Review of survival analyses published in cancer journals
DG Altman, BL Stavola, SB Love and KA Stepniewska
British Journal of Cancer, 1995

- Review of 132 papers analysing survival data
- The papers were published in British Journal of Cancer, European Journal of Cancer, Journal of Clinical Oncology, American Journal of Clinical Oncology and Cancer between October and December 1991
- The review was not restricted to observational epidemiology; keep in mind that praxis differs between disciplines
- After reviewing the papers the authors suggest guidelines for presentation of survival analyses

Publishing your study: describing the data

- Describe how you collected/obtained the data. If data are obtained from a registry, give a brief description of the registry and information on quality and completeness (e.g. reference to Barlow et al. [2]).
- Describe inclusion criteria; e.g. dates of diagnosis, end of follow-up, etc
- Describe exclusion criteria; e.g. how did you handle DCO and autopsy cases
- Report how many subjects were lost to follow-up and how they were handled in the analysis
- Report the sample size (number of individuals ever at risk) and number of events for each end point

Publishing your study: follow-up and end points

- Define the time-scale, what is the time origin and the event(s) of interest
- Define any censoring events
- Define end of study
- Report a summary of the length of follow-up; e.g. median, min, max

Publishing your study: describing the statistical methods

- Name the method used for estimating survival probabilities
- Name any test used in the analysis
- What regression model did you use (how was follow-up time modeled)
- Report all covariates included in analysis, why they were included and how they were modeled (categorical, linear, spline)
- Comment on missingness and how it was handled in the analysis
- Report test of proportional hazards, and if the PH assumption did not hold how was that handled
- Name the software used

Publishing your study: presentation of results

- Give a summary of overall survival, for example, median or percent surviving *n* years
- If relevant, also give a summary of survival for different groups
- When presenting results of a logrank (or similar) test, report the p-value and also report the numbers of observed and expected events in each group
- When presenting results from a regression analysis, report the estimated hazard ratios with their confidence intervals and p-values[1]

[1] Standard practice in epidemiology is to report only HR and CI

Publishing your study: graphs

- Use meaningful time intervals and label the axes appropriately
- Consider marking survival time of censored observations and to give number at risk at selected time points (see the CRAN Survival Taskview for packages)
- If several curves are reported in the same plot use different line types
- Mark confidence intervals
- (comment from Paul) It's traditional to present graphs $S(t)$ (e.g. Kaplan-Meier) as a simple descriptive analysis but then model the rates. If you have no particular interest in $S(t)$, consider presenting graphs of the hazard instead.

Course review I

Forms of data Right censoring implies that the event of interest happens after the last time. Left truncation means that we condition on no event to the entry time. We can measure time with different time origins (e.g. birth, diagnosis, treatment, calendar period).

Course review I

Forms of data Right censoring implies that the event of interest happens after the last time. Left truncation means that we condition on no event to the entry time. We can measure time with different time origins (e.g. birth, diagnosis, treatment, calendar period).

Describe your data For right-censored data, use Kaplan-Meier curves. We could also smooth the rates whilst interpreting the results cautiously.

Course review II

Comparing two or more groups Many options!

- Cox regression, with a non-parametric baseline hazard. Estimands include hazard ratios and median survival.
- Flexible parametric models (FPMs). Estimands include hazard ratios, hazard differences, survival differences, and differences in restricted mean survival times.
- Poisson regression. Estimands include rate ratios, rate differences and survival differences.
- Accelerated failure time models, with acceleration factors (and hazard differences, etc).
- Additive hazards models, with differences in cumulative hazards.

Course review III

Time-varying hazard ratios Again, many options.

- Cox regression, using time-splitting or `tt()` functions. Test for difference and plot using Schoenfeld residuals. We can also use a stratified Cox model (if the strata are essentially nuisance terms).
- Poisson regression, using time-splitting and including time as another covariate, with interactions between time and covariates. Also allows for multiple time scales.
- FPMs using time-splitting or using splines.

Course review III

Time-varying hazard ratios Again, many options.

- Cox regression, using time-splitting or `tt()` functions. Test for difference and plot using Schoenfeld residuals. We can also use a stratified Cox model (if the strata are essentially nuisance terms).
- Poisson regression, using time-splitting and including time as another covariate, with interactions between time and covariates. Also allows for multiple time scales.
- FPMs using time-splitting or using splines.

Study designs Our focus has been on cohort studies and randomised controlled trials. Others:

- Nested case-control studies based on risk-set sampling (closely related to Cox regression).
- Case-cohort studies: not described (requires weights, which complicates the analysis a little).
- Matched cohort studies: not described (typically we adjust for the matching factors).

Course review IV

Testing

- Be clear about your null hypothesis
- A likelihood ratio test is preferred over a Wald test
- Be careful with sub-group analyses:)

Course review IV

Testing

- Be clear about your null hypothesis
- A likelihood ratio test is preferred over a Wald test
- Be careful with sub-group analyses:)

Model building

- For causal modelling, we propose using a pre-specified analysis plan with a causal diagram to address which potential confounders should be adjusted for.
- Sometimes we need to choose which model to use for adjustment (e.g. degrees of freedom for FPMs). Make your analysis plan your friend.

Estimands and estimators

Estimand	Estimator	Notes
Survival	Kaplan-Meier	Non-parametric; small bias
	Poisson regression	Awkward post-estimation
	Cox model + Breslow	
	Flexible parametric model	
	Aalen's additive hazards	
Cumulative hazard	Nelson-Aalen	Non-parametric
Hazard	Nelson-Aalen + Kernel density	Smoothed
Rate	count/(person-time)	Poisson distribution
	Poisson regression	
Rate ratio	Poisson regression	
Hazard ratio	Cox model	Non-parametric baseline
Time-dependent rate ratio	Poisson model	
Time-dependent hazard ratio	Cox model	Inflexible implementations
	Flexible parametric model	
Time-dependent additive hazards	Aalen's additive hazards model	Collapsible
Proportional odds	Flexible parametric model	
Probit	Flexible parametric model	
Accelerated failure time	Flexible parametric model	Collapsible

Restricted mean survival time (RMST)

- An increasingly popular measure for time-to-event outcomes is the restricted mean survival time (RMST), which is the area under the survival curve up to a specified time t (e.g. five years) given covariates x. Mathematically:

$$\text{RMST}(t|x) = \int_0^t S(u|x)du$$

- This has some nice properties; see `https://doi.org/10.1186/1471-2288-13-152`.
- We may also be interested in differences in RMSTs.

Advanced topics in survival analysis I

Emulated target trials For causal inference using observational data, we may be interested in emulating a target randomised trial. Hernán and Robins [3] provide an introduction to this important epidemiological method. The target emulation should account for:

- Eligibility criteria (as per the target trial)
- Treatment strategies (pragmatic interventions)
- Assignment procedures (without blinding)
- Follow-up period (including the careful definition of time zero)
- Outcome (validated?)
- Causal contrast of interest
- Analysis plan

Frailty models Individuals within a group may have more similar hazards than individuals between groups. We can model for the variance between groups using frailty models.

Recurrent events Individuals may experience one or more of a type of event.

Advanced topics in survival analysis II

Relative survival If cause of death is not well coded, then we can model the overall mortality hazard $h(t) = h_0(t) + h_e(t)$, where $h_0(t)$ is the background mortality hazard and $h_e(t)$ is the excess mortality hazard. This is useful for comparing survival between cancer registries and for survival predictions outside of observed studies.

Competing risks What is the probability of experiencing a particular event in a defined period in the presence of competing events? For example, are men more likely than women to be admitted to hospital between ages 70-79 years for stroke? Men may have a higher incidence of stroke – but they are also more likely to die due to other causes. In this setting, do we not censor for the competing events.

Multi-state models To generalise competing risks, an individual may move between different health/disease states, with competing events. We can then estimate the proportion of individuals in each of the states at a given time. Several courses have been held on competing risks and multi-state models (Putter and Geskus; Crowther and Lambert).

Exercises for Monday afternoon

22. Estimating the effect of a time-varying exposure – the bereavement data
23. Estimating SMRs
25. Localised melanoma: Generating and analysing a nested case-control study
28. Model cause-specific survival using flexible parametric models [This is a key exercise]

References I

Odd O. Aalen, Richard J. Cook, and Kjetil Røysland.
Does Cox analysis of a randomized survival study yield a causal treatment effect?
Lifetime Data Analysis, 21(4):579–593, Oct 2015.

Lotti Barlow, Kerstin Westergren, Lars Holmberg, and Mats Talback.
The completeness of the Swedish Cancer Register - a sample survey for year 1998.
Acta Oncol, pages 1–7, Sep 2008.

Miguel A. Hernán and James M. Robins.
Using big data to emulate a target trial when a randomized trial is not available.
American Journal of Epidemiology, 183(8):758–764, 2016.

Britta Krynitz, Gustaf Edgren, Bernt Lindelöf, Eva Baecklund, Christina Brattström, Henryk Wilczek, and Karin E. Smedby.
Risk of skin cancer and other malignancies in kidney, liver, heart and lung transplant recipients 1970 to 2008–a Swedish population-based study.
Int J Cancer, 132(6):1429–1438, Mar 2013.

Subgroup analyses are evil – use an interaction model

- Assume that we have a group or stratification variable k and that our research question is whether the hazard ratio for an exposure x varies by k.
- One approach is to do a sub-group analysis, where we do a separate model for each level of k:
  - We would separately adjust for other covariates in each of the models
  - For each model, our null hypothesis is that the adjusted hazard ratio for x is 1
- An alternative approach is to use an interaction model with main effects and x and k and interactions between x and k:
  - Our null hypothesis is that the adjusted hazard ratios for x for each level of k is the same
- The null hypotheses for the two approaches are different. The second approach addresses our research question.
- All else being equal, a larger sub-group is more likely to detect a significant effect. For such a group, that does not imply a significantly higher effect in that group compared with the other groups.
- In summary, avoid sub-group analyses – or interpret them very carefully.