Biostatistics III:
Survival analysis for epidemiologists in R

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

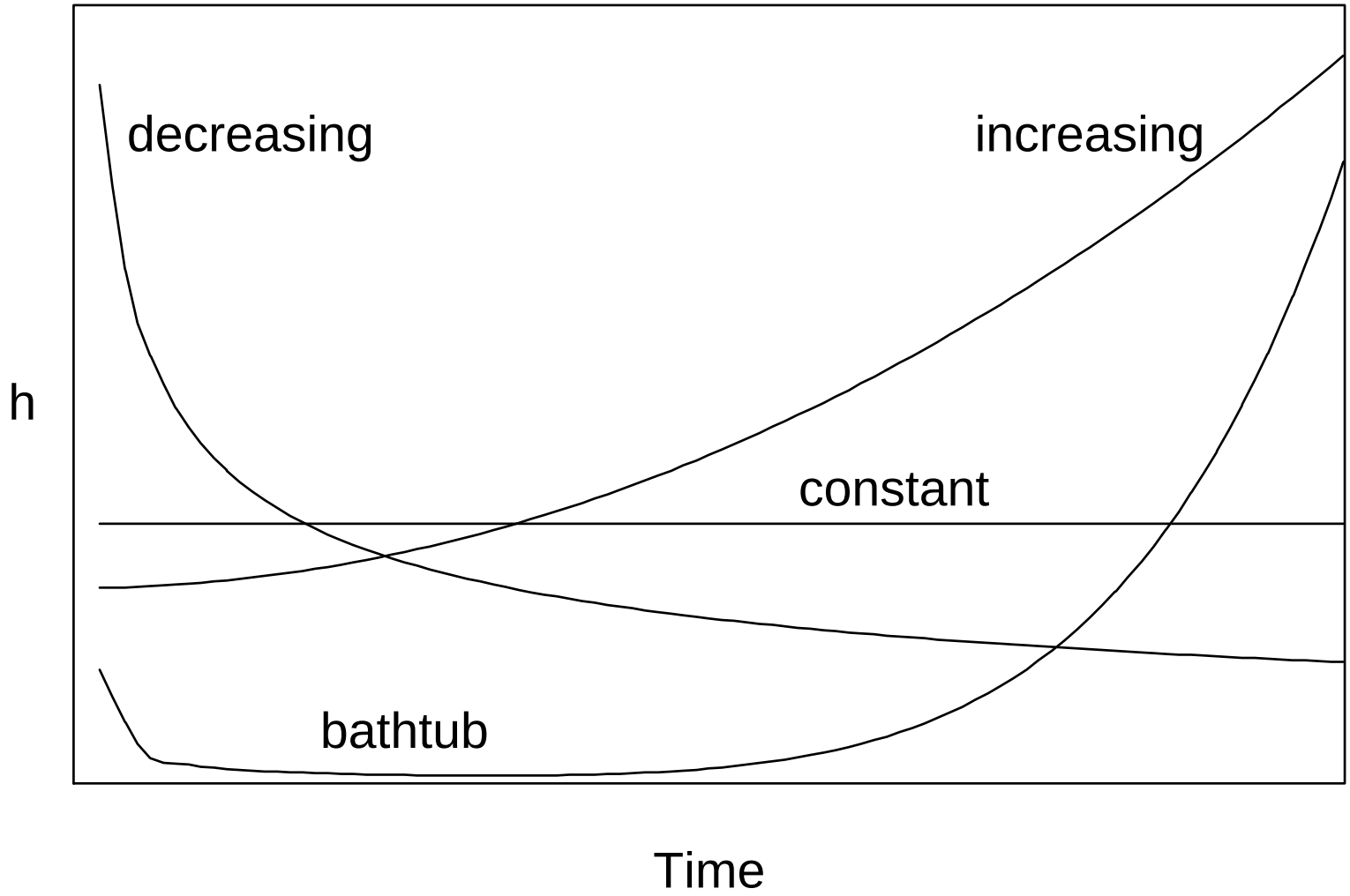
6–15 November, 2023

http://kiwas.ki.se/katalog/katalog/kurs/9735, course 2992

Topics for Day 3

- Rates which vary over time
- The Cox model
- Comparison of Cox and Poisson regression
- Assessing the proportional hazards assumption in Cox regression
- Modelling time-varying rates and exposures

Common forms for the hazard function I



Common forms for the hazard function II

- A bathtub-shaped hazard is appropriate for all cause mortality in most human populations followed from birth.
- A decreasing hazard function is appropriate following the diagnosis of most types of cancer, where mortality due to the cancer is highest immediately following diagnosis, and then decreases with time.
- A constant hazard function is often used for modelling the lifetime of electronic components, but is also appropriate following the diagnosis of some types of cancer.
- A constant hazard function implies that survival times can be described by an exponential distribution (which has one parameter, the hazard λ). This distribution is 'memoryless' in that the expected survival time for any individual is independent of how long the individual has survived so far.
- The survivor function has the same basic shape (a nonincreasing function from 1 to 0) for all types of data and the hazard function is often a more informative means of studying differences between patient groups.

Shape of the hazard in Poisson regression

- The Poisson regression model is

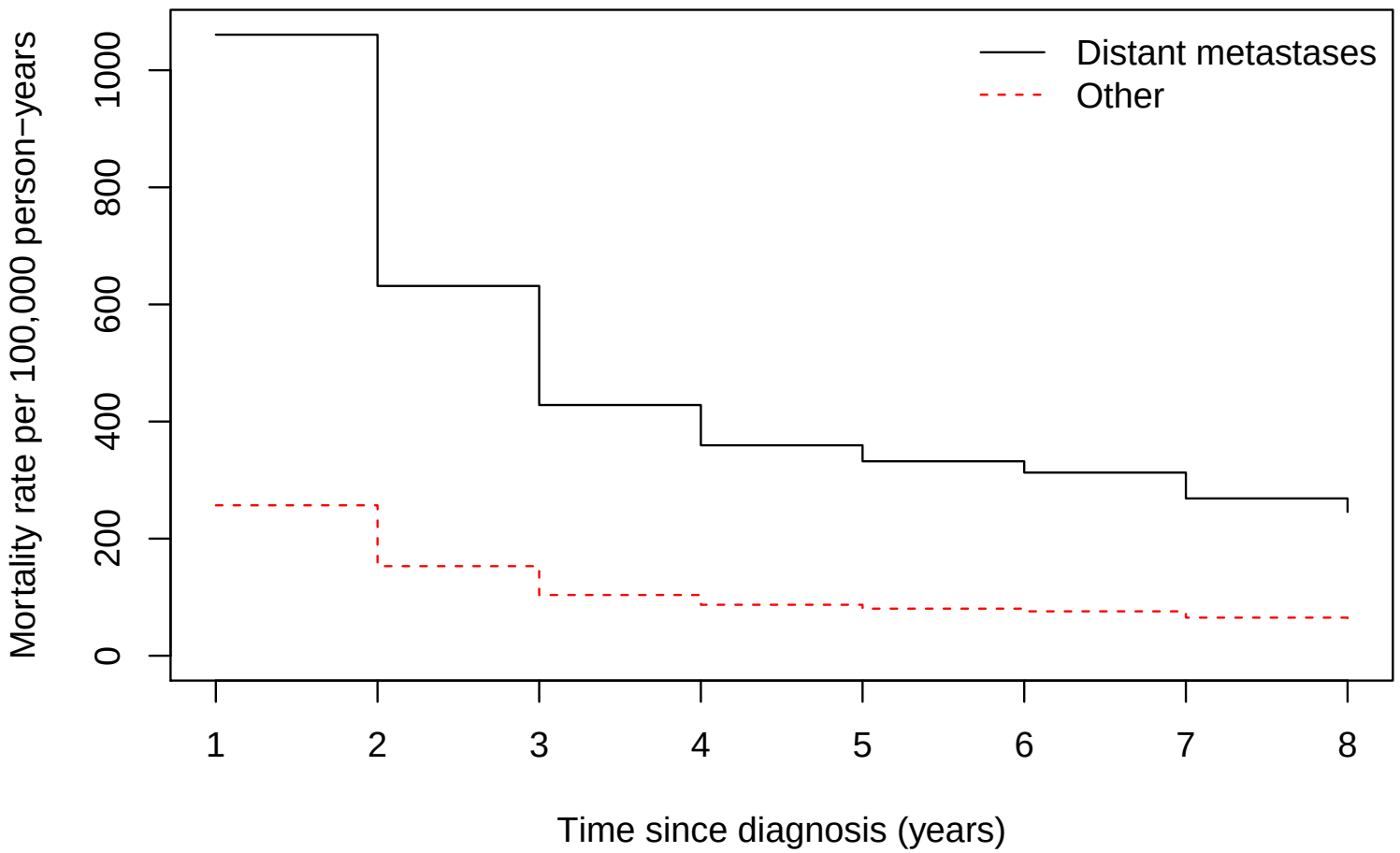
$$\begin{aligned} \log(\lambda) &= \beta_0 + \beta_1 X \\ \lambda &= \exp(\beta_0 + \beta_1 X) \\ &= \exp(\beta_0)\exp(\beta_1 X) \end{aligned}$$

- The baseline hazard is constant in a Poisson regression, $\exp(\beta_0)$.
- If we add a categorical variable for time, e.g. time-since-entry in 1-year bands, then the baseline hazard is a step function of time. The hazard is piecewise constant in 1-year bands.

$$\lambda = \exp(\beta_0 + \beta_3 t_{[1,2)} + \beta_4 t_{[2,3)} + \cdots)\exp(\beta_1 X)$$

where $t_{[1,2)}$ is an indicator for time being in the interval $[1, 2)$. Note that $t_{[0,1)}$ if left out from the equation (it is assumed to be the reference time band)

Shape of hazard: step function



Parametric models I

- If we assume that survival times follow an exponential distribution, we could model the hazard as a function of one or more covariates.
- We could then obtain an estimate of the hazard ratio for the treatment group compared to the control group while adjusting for other explanatory variables.
- The disadvantage of this method is that assuming an exponential distribution for survival times implies the assumption of a constant hazard function over time, which may not be appropriate.
- We have several options:
  1. Split by time and assume that the hazard is piece-wise constant (Poisson regression – discussed on Day 2).
  2. Use a more flexible distribution (e.g. Weibull regression and flexible parametric survival models).
  3. Use a non-parametric function for the baseline hazard (Cox regression).

Parametric models II

- The Weibull, log-normal and Gompertz distributions have proved to be applicable in several types of medical survival studies. These models can fitted as accelerated failure time (AFT) models. If the time to event is represented by T, with covariates X, then the AFT model assumes that

$$E(T|X) = E(T_0)exp(-\beta X)$$

  where T_0 is the baseline time variable for $X = 0$.
- These models have a very nice interpretation, where $\exp(-\beta)$ is the change in the mean time per unit change in X. As we will discuss on Day 4, these models are also collapsible (the estimated effect of an exposure does not change if we add or remove explanatory variables that are *not* confounders).
- If a parametric distribution is appropriate, such models will result in more efficient estimates (narrower confidence limits) of the parameters of interest.
- An alternative parametric class are the flexible parametric survival models (or generalised survival models), which includes proportional hazards and proportional odds models. We will also discuss this model class further on Day 4.

Parametric models III

- The parametric models may be sensitive to the choice of parametric distribution.
- That is, when using parametric survival models, special attention must be paid to testing the appropriateness of the model.

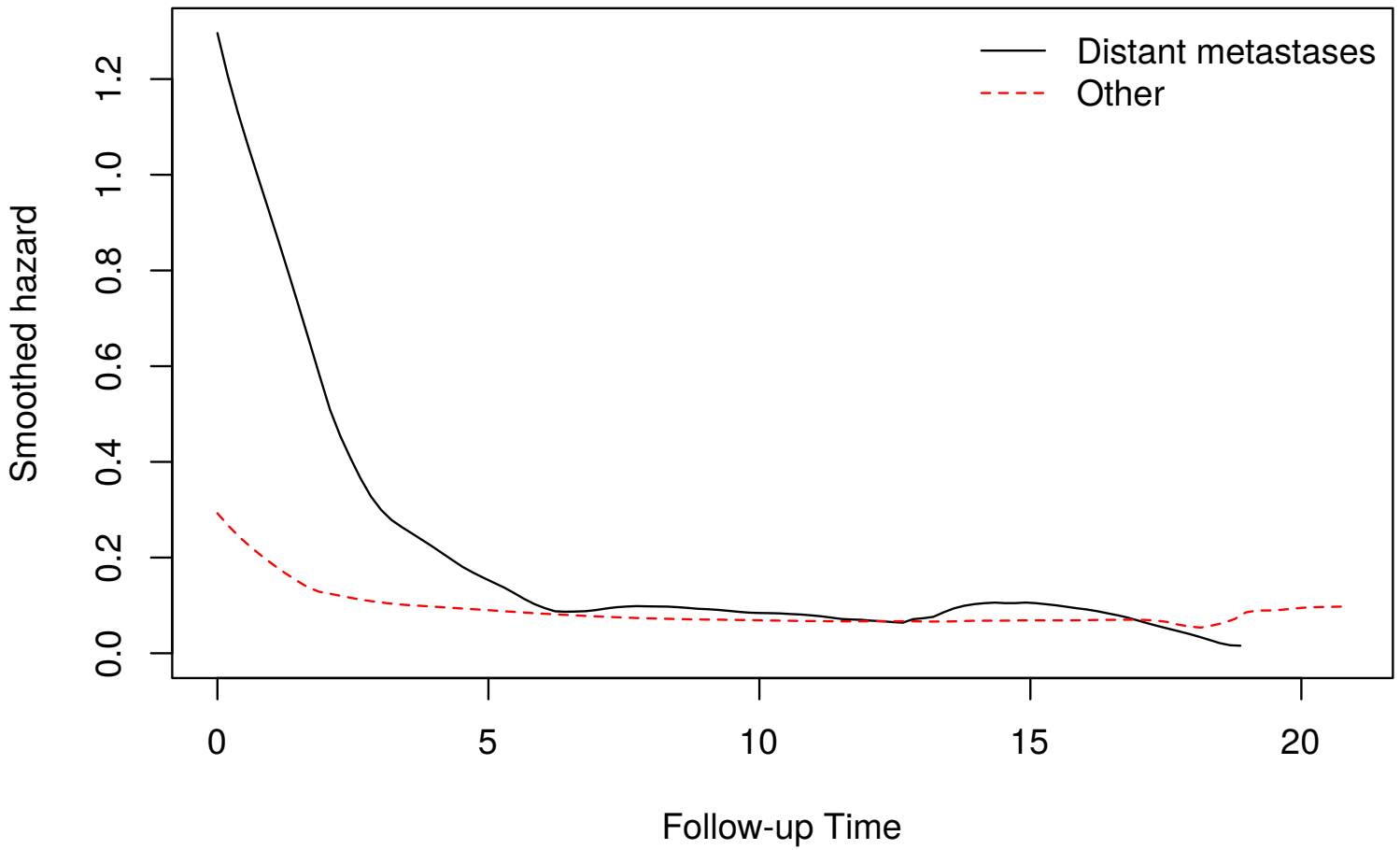
The shape of hazards in a Cox model

- The Cox model does not make any assumption about the shape of the hazard function.
- Instead, the baseline hazard is allowed to vary freely.
- The Cox model only estimates hazard ratios relative to the baseline hazard.
- Since the baseline hazard is not estimated in the Cox model, it is said to be semi-parametric.
- Note: given a fitted Cox model, we can estimate the non-parametric baseline cumulative hazard and then smooth to estimate the hazard.

An introduction to the Cox model via an example: Survival of patients diagnosed with colon carcinoma I

- Patients diagnosed with colon carcinoma 1984–95. Potential follow-up to end of 1995; censored after 10 years.
- Outcome is death due to colon carcinoma.
- Interest is in the effect of clinical stage at diagnosis (distant metastases vs no distant metastases).
- How might we specify a statistical model for these data?

An introduction to the Cox model via an example: Survival of patients diagnosed with colon carcinoma II



The Cox proportional hazards model I

- A proportional hazards model is on the form

$$\lambda(t|X) = \lambda_0(t)\exp(\beta X).$$

- The hazard at time t for an individual with some covariate values is a multiple of the baseline.
- This means that the hazards for different levels of X are proportional.
- The Cox model is a proportional hazards model.
- However, the Cox model does not estimate the baseline hazard, $\lambda_0(t)$. It only estimates the regression coefficients, β.
- The 'intercept' in the Cox model [3], the hazard (event rate) for individuals with all covariates X at the reference level, is an arbitrary function of time[1], often called the baseline hazard and denoted by $\lambda_0(t)$.

The Cox proportional hazards model II

- The Cox model can also be written on the log scale

$$\log[\lambda(t|X)] = \log[\lambda_0(t)] + \beta X.$$

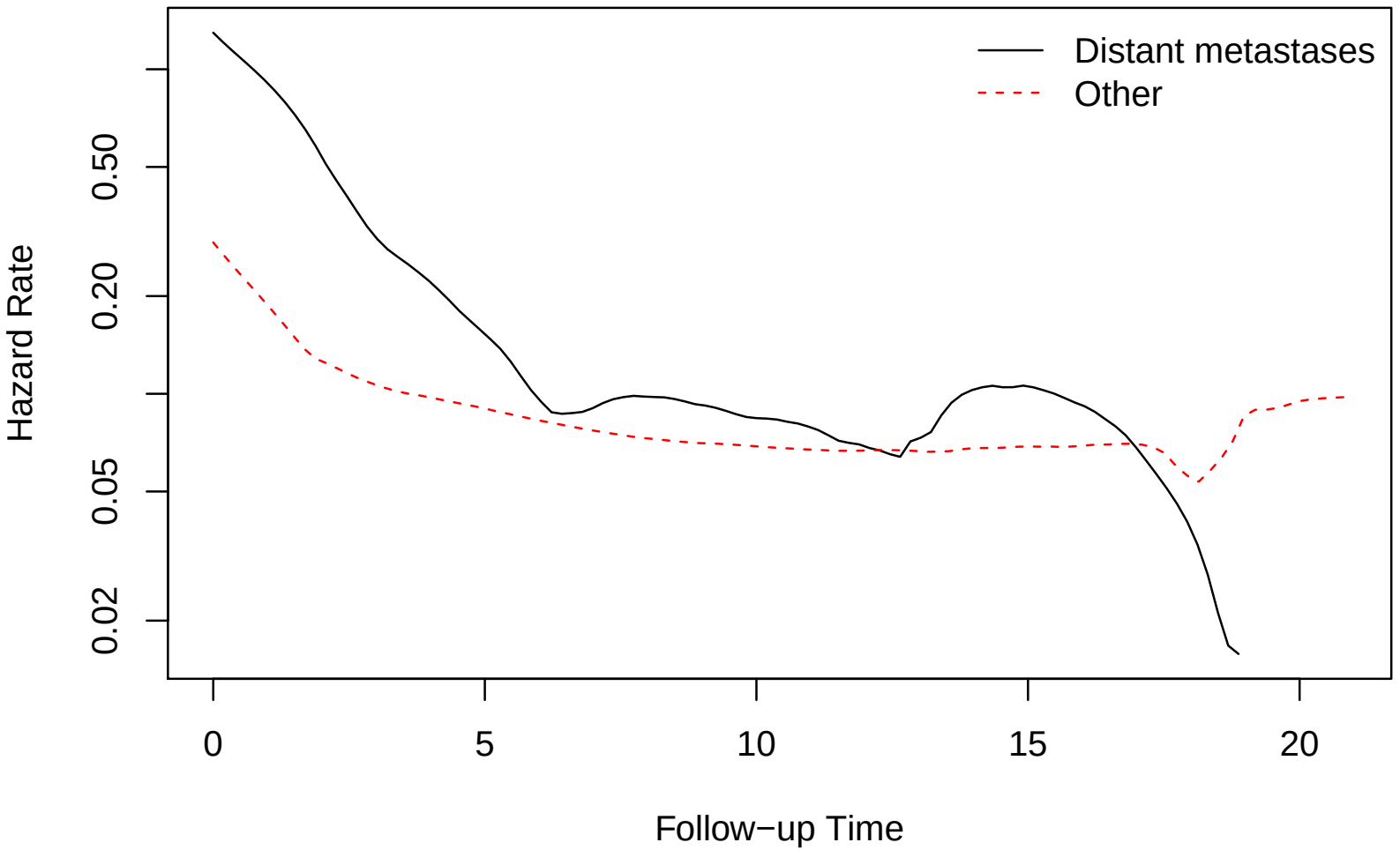
  where $X = 1$ for patients with distant metastases at diagnosis and $X = 0$ for patients without distant metastases at diagnosis.

- The difference between two hazards is a constant β regardless of t

$$\log[\lambda(t|X)] - \log[\lambda_0(t)] = \beta X.$$

- The two hazard curves are thus assumed to be parallel on a log scale.

The Cox proportional hazards model III



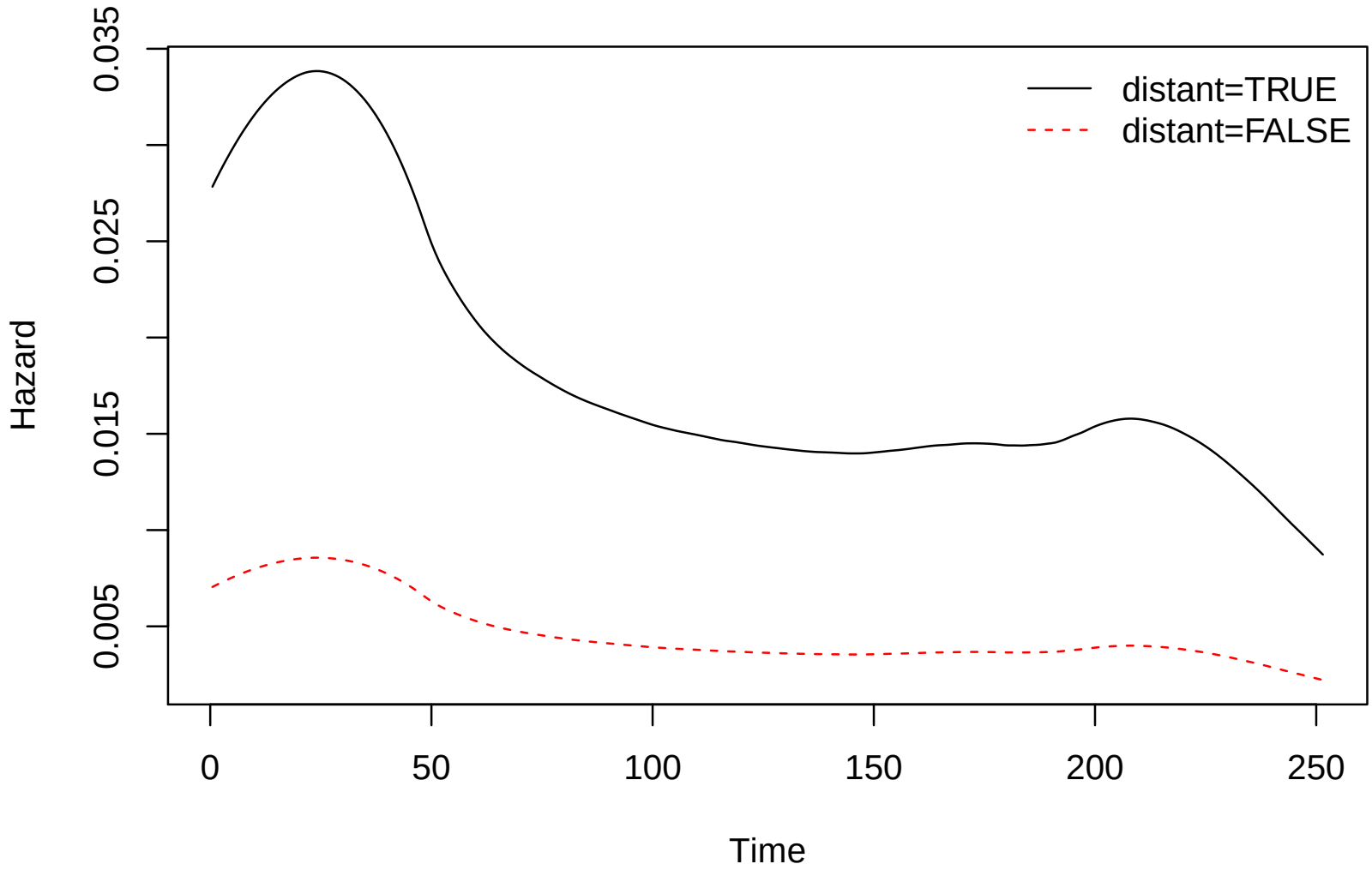
[1] time t can be defined in many ways, e.g., attained age, time-on-study, calendar time, etc.

Fit a Cox model to estimate the mortality rate ratio I

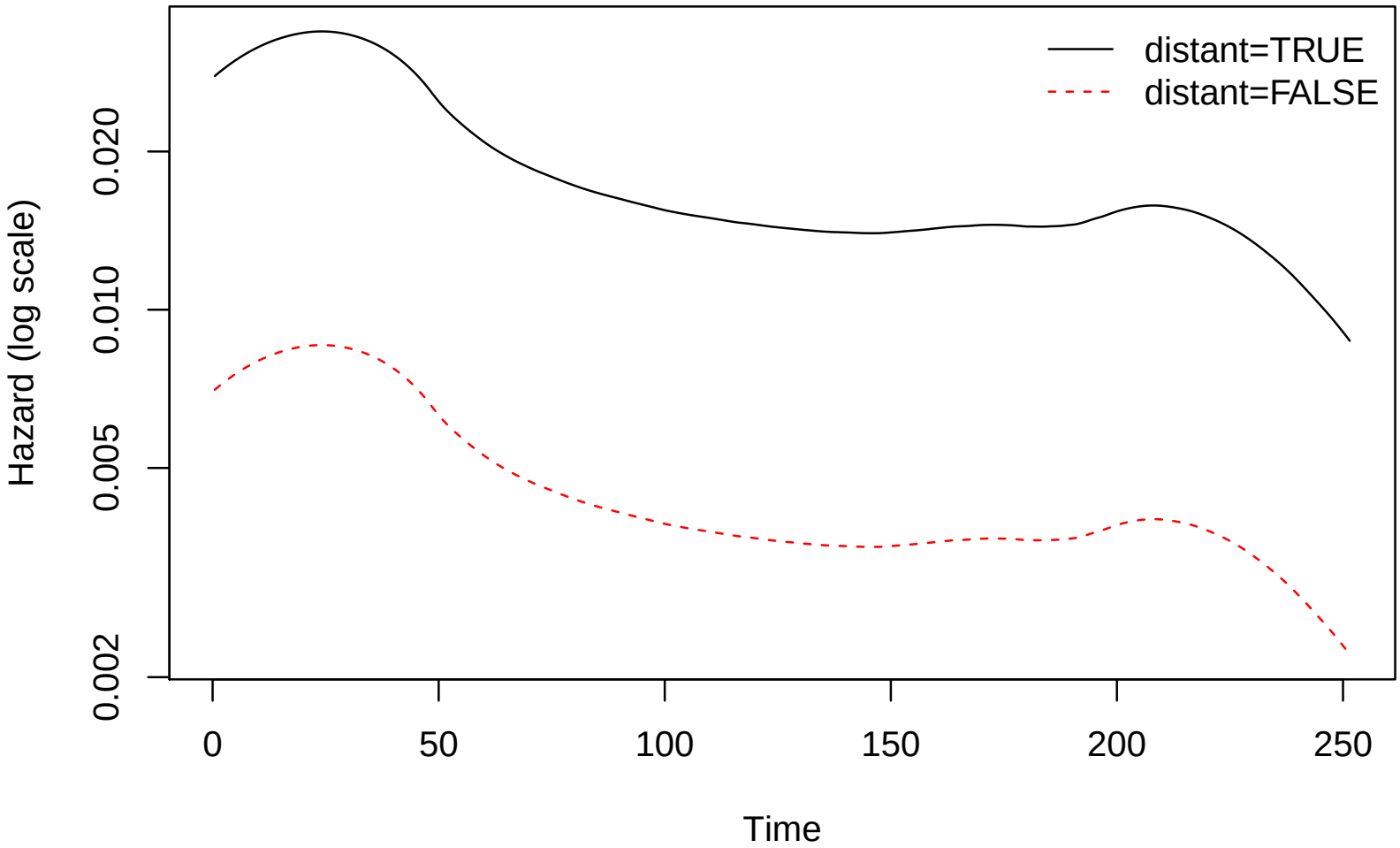
R code and output

```
> colon2 <- transform(colon, distant = (stage == "Distant"),
                      dead = status %in% c("Dead: cancer","Dead: other"))
> summary(coxph(Surv(surv_mm,dead)~distant, data=colon2))

Call:
coxph(formula = Surv(surv_mm, dead) ~ distant, data = colon2)

  n= 15564, number of events= 10918

               coef exp(coef) se(coef)     z Pr(>|z|)
distantTRUE 1.37395   3.95093  0.02033 67.59   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

            exp(coef) exp(-coef) lower .95 upper .95
distantTRUE     3.951     0.2531     3.797     4.112
```

Fit a Cox model to estimate the mortality rate ratio II



Fit a Cox model to estimate the mortality rate ratio III



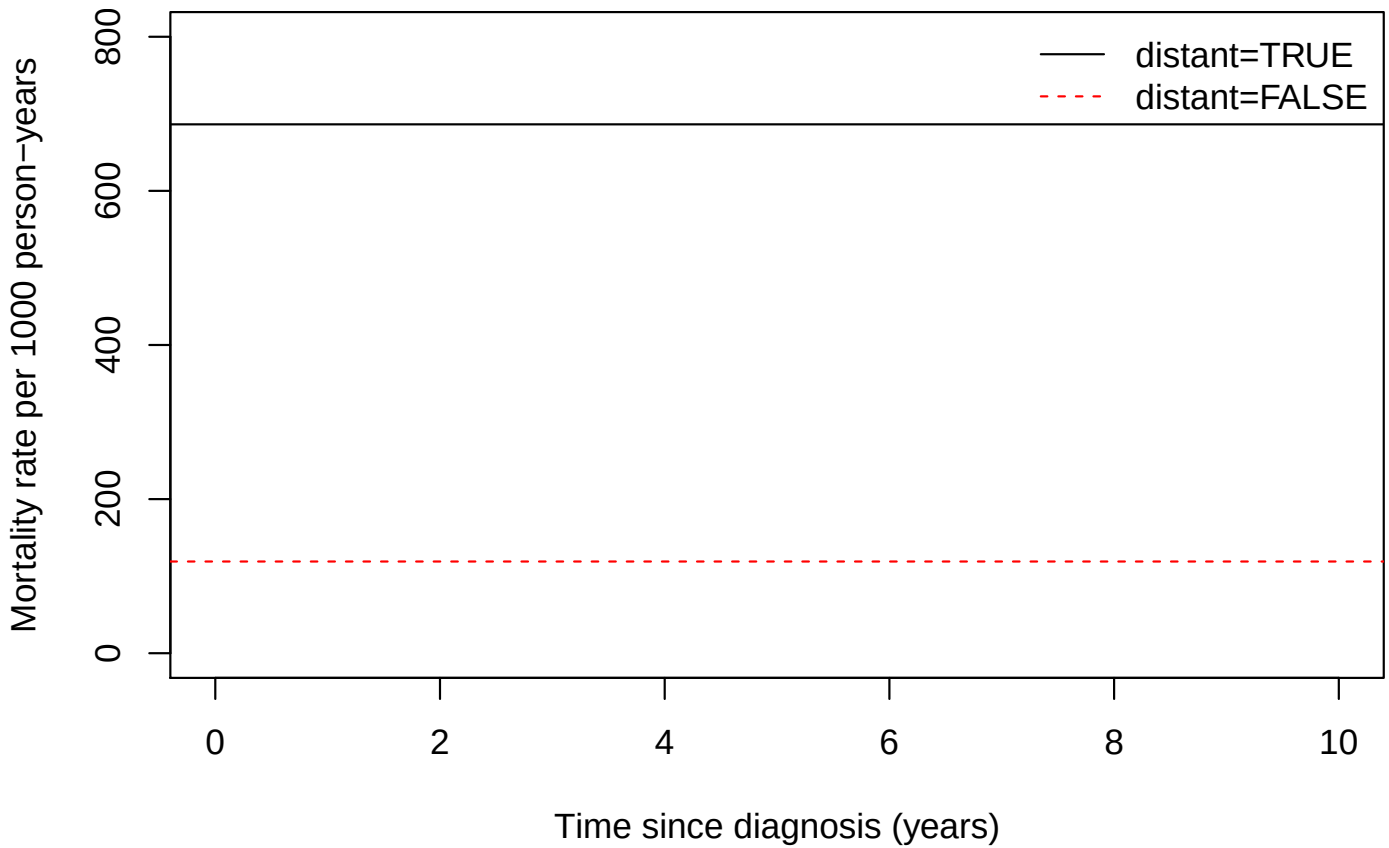
An analogous Poisson regression model?

R code and output

```
> summary(glm(dead ~ distant + offset(log(surv_mm)), data=colon2, family=poisson))

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -4.61317    0.01278 -361.10   <2e-16 ***
distantTRUE  1.75189    0.01928   90.84   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

- Is this conceptually analogous to the Cox model with one predictor (`distant`)?

Fitted values: Poisson model with one predictor (distant)



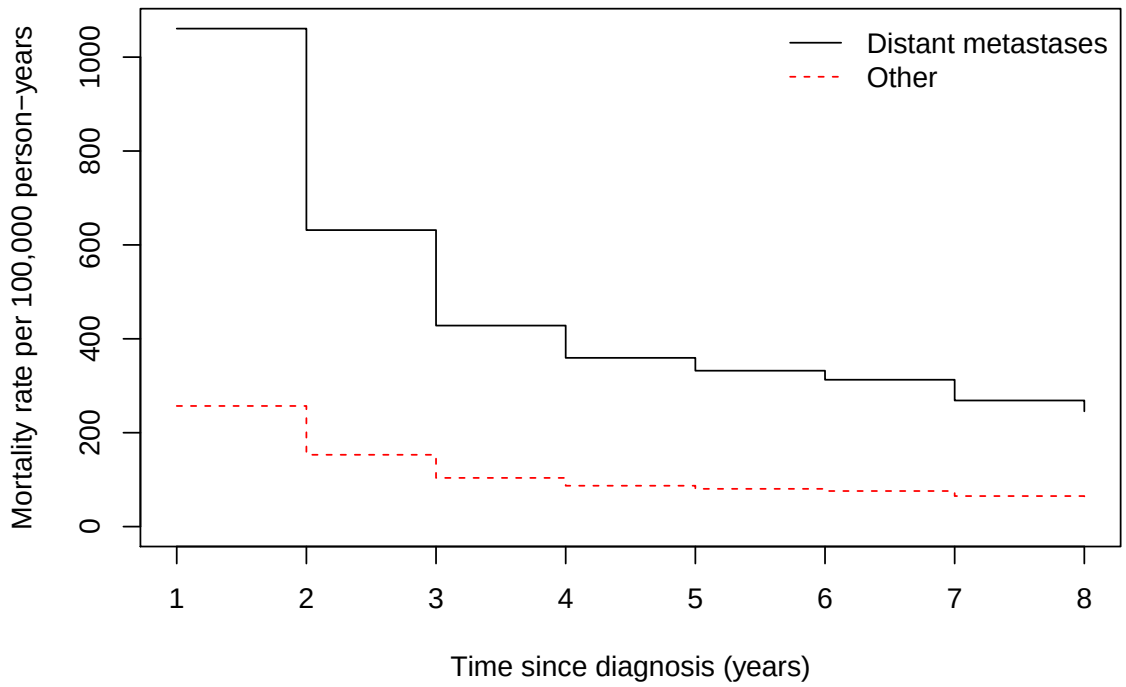
- We haven't controlled for time, whereas the Cox model does.

An analogous Poisson regression model

R code and output

```
> colon3 <- survSplit(Surv(surv_mm/12,  dead) ~ distant, data=colon2, episode="timeband",
                      cut=1:8)
> colon3 <- transform(colon3, py = tstop - tstart)
> fit <- glm(dead ~ distant + factor(timeband) + offset(log(py)), data=colon3,
             family=poisson)
> eform(fit)
                   exp(beta)     2.5 %    97.5 %
(Intercept)        0.2569548 0.2481677 0.2660530
distantTRUE        4.1281424 3.9685036 4.2942028
factor(timeband)2 0.5954304 0.5647893 0.6277339
factor(timeband)3 0.4036984 0.3761544 0.4332593
factor(timeband)4 0.3388328 0.3113618 0.3687275
factor(timeband)5 0.3130855 0.2843180 0.3447637
factor(timeband)6 0.2948831 0.2646559 0.3285626
factor(timeband)7 0.2532605 0.2231677 0.2874112
factor(timeband)8 0.2315131 0.2006489 0.2671249
factor(timeband)9 0.2262279 0.2095410 0.2442436
```

Fitted values: Poisson regression model adjusted for time I



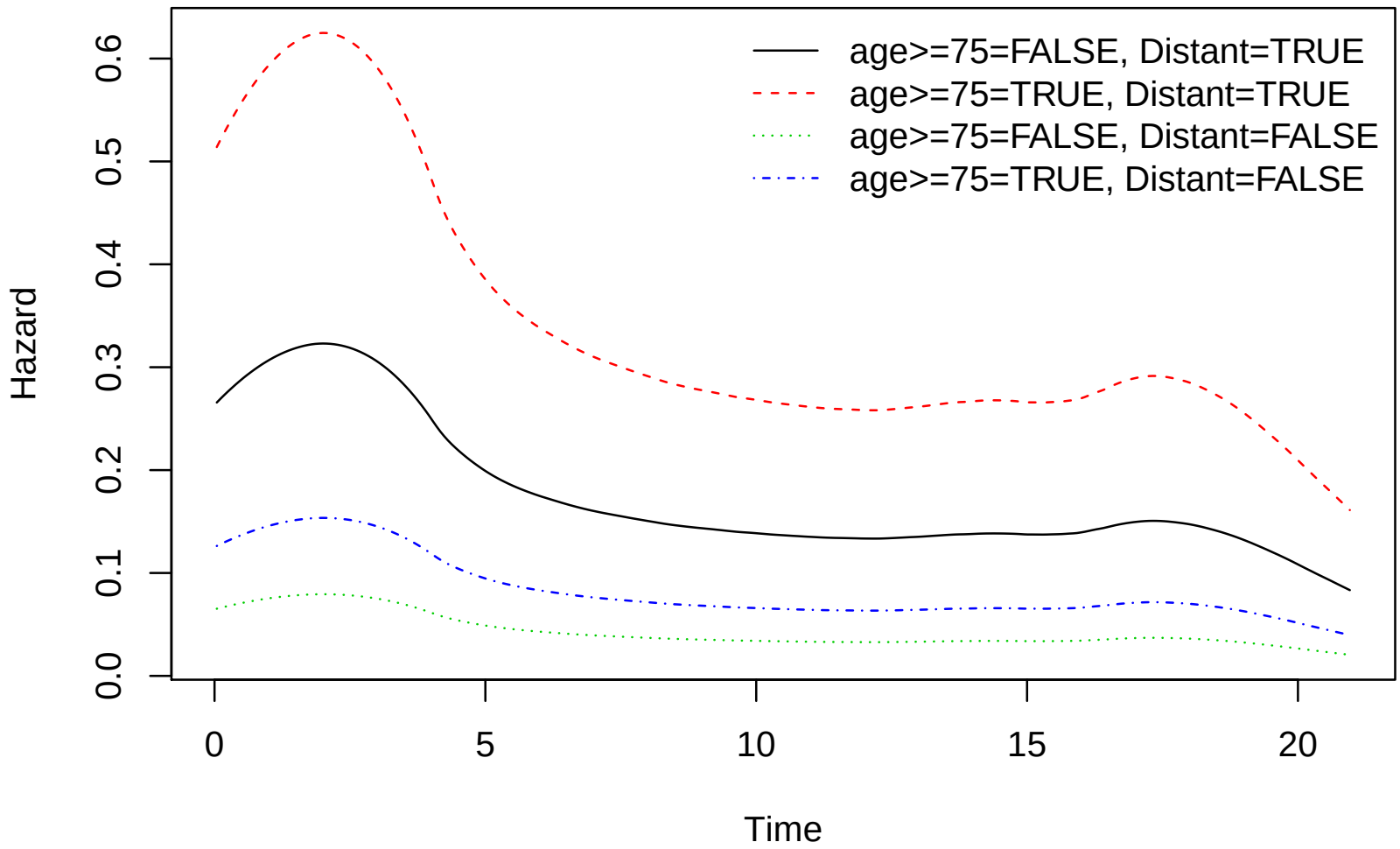
- Once we adjust for time we get a similar estimate for the effect of distant.

Fit a Cox model adjusted for age at diagnosis I

R code and output

```
  fit <- coxph(Surv(surv_mm/12, status=="Dead: cancer") ~ I(age>=75) +
                    I(stage=="Distant"), data=colon)
  summary(fit)

  n= 15564, number of events= 8369

                              coef exp(coef) se(coef)     z Pr(>|z|)
I(age >= 75)TRUE           0.49319   1.63754  0.02232 22.10   <2e-16 ***
I(stage == "Distant")TRUE 1.68755   5.40620  0.02295 73.53   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

                           exp(coef) exp(-coef) lower .95 upper .95
I(age >= 75)TRUE               1.638     0.6107     1.567     1.711
I(stage == "Distant")TRUE      5.406     0.1850     5.168     5.655
```

Fit a Cox model adjusted for age at diagnosis II



The Cox proportional hazards model (in detail) I

- The most commonly applied model in medical time-to-event studies is the Cox proportional hazards model [3].
- The Cox proportional hazards model does not make any assumption about the shape of the underlying hazards, but makes the assumption that the hazards for patient subgroups are proportional over follow-up time.
- We are usually more interested in studying how hazard varies as a function of explanatory variables (the relative hazard) rather than the shape of the underlying hazard function (the absolute hazard).
- In most statistical models in epidemiology (e.g. linear regression, logistic regression, Poisson regression) the outcome variable (or a transformation of the outcome variable) is equated to the 'linear predictor', $\beta_0 + \beta_1 X_1 + \cdots + \beta_k X_k$.
- $X_1, \ldots, X_k$ are explanatory variables and $\beta_0, \ldots, \beta_k$ are regression coefficients (parameters) to be estimated.
- The Xs can be continuous (age, blood pressure, etc.) or if we have categorical predictor variables we can create a series of indicator variables (Xs with values 1 or 0) to represent each category.

The Cox proportional hazards model (in detail) II

- A common choice is

$$\lambda(t|\boldsymbol{X}) = \exp(\beta_0 + \beta_1 X_1 + \cdots + \beta_k X_k)$$
$$\log \lambda(t|\boldsymbol{X}) = \beta_0 + \beta_1 X_1 + \cdots + \beta_k X_k$$

- This formulation is identical to the Poisson regression model. That is,

$$\log\left(\frac{\text{no. events}}{\text{person-time}}\right) = \beta_0 + \beta_1 X_1 + \cdots + \beta_k X_k$$

- The one flaw in this potential model is that $\lambda(t|\boldsymbol{X})$ is a function of t, whereas the right hand side will have a constant value once the values of the βs and Xs are known.
- This does not cause any mathematical problems, although experience has shown that a constant hazard rate is unrealistic in most practical situations.
- The remedy is to replace β_0, the 'intercept' in the linear predictor, by an arbitrary function of time — say $\log \lambda_0(t)$; thus, the resulting model equation is

$$\log \lambda(t|\boldsymbol{X}) = \log \lambda_0(t) + \beta_1 X_1 + \cdots + \beta_k X_k.$$

The Cox proportional hazards model (in detail) III

- The arbitrary function, $\lambda_0(t)$, is evidently equal to the hazard rate, $\lambda(t|\boldsymbol{X})$, when the value of $\boldsymbol{X}$ is zero, i.e., when $X_1 = \cdots = X_k = 0$.
- The model is often written as

$$\lambda(t|\boldsymbol{X}) = \lambda_0(t)\exp(\boldsymbol{X}\beta).$$

- It is not important that an individual having all values of the explanatory variables equal to zero be realistic; rather, $\lambda_0(t)$ represents a reference point that depends on time, just as β_0 denotes an arbitrary reference point in other types of regression models.

The Cox proportional hazards model (in detail) IV

- This regression model for the hazard rate was first introduced by Cox [3], and is frequently referred to as the Cox regression model, the Cox proportional hazards model, or simply the Cox model.
- Estimates of $\beta_1, \ldots, \beta_k$ are obtained using the method of maximum partial likelihood.
- As in all other regression models, if a particular regression coefficient, say β_j, is zero, then the corresponding explanatory variable, X_j, is not associated with the hazard rate of the response of interest; in that case, we may wish to omit X_j from any final model for the observed data.
- As with logistic regression and Poisson regression, the statistical significance of explanatory variables is assessed using Wald tests or, preferably, likelihood ratio tests.
- The Wald test is an approximation to the likelihood ratio test. The likelihood is approximated by a quadratic function, an approximation which is generally quite good when the model fits.
- In most situations, the test statistics will be similar.

The Cox proportional hazards model (in detail) V

- Differences between these three test statistics are indicative of possible problems with the fit of the model.
- Because of the inter-relationship between the hazard function, $\lambda(t|X)$, and the survivor function, $S(t|X)$, we can show that the PH regression model is equivalent to specifying that

$$\begin{aligned} S(t|\boldsymbol{X}) &= \exp\left(-\int_0^t \lambda_0(u)\exp(\beta_1 X_1 + \cdots + \beta_k X_k)du\right) \\ &= \exp\left(-\exp(\beta_1 X_1 + \cdots + \beta_k X_k)\int_0^t \lambda_0(u)du\right) \\ &= \exp\left(-\int_0^t \lambda_0(u)du\right)^{\exp(\beta_1 X_1 + \cdots + \beta_k X_k)} \\ &= S_0(t)^{\exp(\beta_1 X_1 + \cdots + \beta_k X_k)} \end{aligned} \tag{1}$$

where $S(t|\boldsymbol{X})$ denotes the survivor function for a subject with explanatory variables $\boldsymbol{X}$, and $S_0(t)$ is the corresponding survivor function for an individual with all covariate values equal to zero.

The Cox proportional hazards model (in detail) VI

- The assumption of proportional hazards is a strong assumption, and should be tested (see slide 55).
- Most software packages, will provide estimates of $S(t)$ based on the fitted proportional hazards model for any specified values of explanatory variables.
- For example, the `biostat3::coxphHaz` function can be used to plot the hazard function.

The Estimated Regression Coefficients

- The estimated coefficients, β, are log rate ratios. To get the rate ratios we need to exponentiate the coeffecients, exp(β).
- The confidence intervals for the β are on the log scale. The CIs are therefore not symmetric around the rate ratios.

Interpreting the Estimated Regression Coefficients I

- Recall that the basic proportional hazard (PH) regression model specifies

$$\lambda(t|\boldsymbol{X}) = \lambda_0(t)\exp(\beta_1 X_1 + \cdots + \beta_k X_k)$$

  equivalently,

$$\log \lambda(t|\boldsymbol{X}) = \log \lambda_0(t) + \beta_1 X_1 + \cdots + \beta_k X_k$$

- Note the similarity to the basic equation for multiple linear regression, i.e.,

$$Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_k X_k$$

- In ordinary regression we derive estimates of all the regression coefficients, i.e., $\beta_1, \ldots, \beta_k$ and β_0.
- In PH regression, the baseline hazard component, $\lambda_0(t)$, vanishes from the partial likelihood; we only obtain estimates of the regression coefficients associated with the explanatory variates $X_1, \ldots, X_k$.

Interpreting the Estimated Regression Coefficients II

- Consider the simplest possible setup, one involving only a single binary variable, X; then the PH regression model is

$$\log \lambda(t|X) = \log \lambda_0(t) + \beta X$$

or equivalently,

$$\begin{aligned} \beta X &= \log \lambda(t|X) - \log \lambda_0(t) \\ &= \log \left(\frac{\lambda(t|X)}{\lambda_0(t)} \right) \end{aligned} \tag{2}$$

- Since $\lambda_0(t)$ corresponds to the value $X = 0$,

$$\beta = \log \left(\frac{\lambda(t|X=1)}{\lambda_0(t)} \right) \tag{3}$$

Interpreting the Estimated Regression Coefficients III

- That is, β is the logarithm of the ratio of the hazard rate for subjects belonging to the group denoted by $X = 1$ to the hazard function for subjects belonging to the group indicated by $X = 0$.
- The parameter β is a log relative rate and $\exp(\beta)$ is a relative rate of response.
- If we conclude that the data provide reasonable evidence to contradict the hypothesis that X is unrelated to response, $\exp(\hat{\beta})$ is a point estimate of the rate at which response occurs in the group denoted by $X = 1$ relative to the rate at which response occurs at the same time in the group denoted by $X = 0$.
- A confidence interval for β, given by $\hat{\beta} \pm 1.96\text{SE}$, represents a range of plausible values for the log relative rate associated with the corresponding explanatory variable.

Interpreting the Estimated Regression Coefficients IV

- Corresponding confidence intervals for the relative rate associated with the same covariate are obtained by transforming the confidence interval for β, i.e.,

$$(\beta_\ell, \beta_u) \Rightarrow \left(e^{\beta_\ell}, e^{\beta_u}\right).$$

- When more than one covariate is involved, the principle is the same; $\exp(\hat{\beta}_j)$ is the estimated relative rate of failure for subjects that differ only with respect to the covariate X_j.
- If X_j is binary, $\exp(\hat{\beta}_j)$ estimates the increased/reduced rate of response for subjects corresponding to $X_j = 1$ versus those denoted by $X_j = 0$.
- When X_j is a numerical (continuous) measurement then $\exp(\hat{\beta}_j)$ represents the estimated change in relative rate associated with a unit change in X_j.
- Since the estimates $\hat{\beta}_1, \ldots, \hat{\beta}_k$ are obtained simultaneously, these estimated relative rates adjust for the effect of all the remaining covariates included in the fitted model.

Example: Localised colon carcinoma 1975–1994 I

- We fitted a proportional hazards model to study the effect of sex, age (in 4 categories), and calendar period (2 categories) on cause-specific mortality (only deaths due to colon cancer were considered events).
- We'll begin by restricting the data to localised cases only (stage=1).
- We consider cause specific mortality.

Example: Localised colon carcinoma 1975–1994 II

R code and output

```
> fit <- coxph(Surv(surv_mm/12, status=="Dead: cancer") ~ sex + agegrp + year8594,
               subset=(stage=="Localised"), data=colon)
> summary(fit)

  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

                         exp(coef) exp(-coef) lower .95 upper .95
sexFemale                   0.9145     1.0935    0.8301    1.0074
agegrp45-59                 0.9493     1.0534    0.7237    1.2453
agegrp60-74                 1.3396     0.7465    1.0470    1.7140
agegrp75+                   2.2572     0.4430    1.7631    2.8900
year8594Diagnosed 85-94     0.7539     1.3265    0.6843    0.8305

Likelihood ratio test= 199.1  on 5 df,   p=0
Wald test            = 198.4  on 5 df,   p=0
```

Example: Localised colon carcinoma 1975–1994 III

- The output commences with a description of the outcome and censoring variable and a summary of the number of subjects and number of failures.
- The default method for handling ties (the Efron method) is used.
- The test statistic `LR chi2(5) = 199.1` is not especially informative. The interpretation is that the 5 parameters in the model (as a group) are statistically significantly associated with the outcome ($P < 0.00005$).
- The factor variable sex is coded 'Male' (the reference) and 'Female'. The estimated hazard ratio for sex represents the ratio of the hazards for females compared to males.
- That is, the estimated hazard ratio is 0.92 indicating that females have an estimated 8% lower colon cancer mortality than males. There is some evidence that the difference is different from zero ($P = 0.07$).
- The model assumes that the estimated hazard ratio of 0.92 is the same at each and every point during follow-up and for all combinations of the other covariates.

Example: Localised colon carcinoma 1975–1994 IV

- That is, the hazard ratio is the same for females diagnosed in 1975–1984 aged 0–44 (compared to males diagnosed in 1975–1984 aged 0–44) as it is for females diagnosed in 1985–1994 aged 75+ (compared to males diagnosed in 1985–1994 aged 75+).
- The factor variable `year8594` is coded 'Diagnosed 75-84' (the reference) and 'Diagnosed 85-94'.
- The estimated hazard ratio is 0.75. We estimate that, after controlling for the time scale, age and sex, patients diagnosed 1985–1994 have a 25% lower mortality than patients diagnosed during 1975–1984. The difference is statistically significant ($P < 0.0005$).
- We chose to group age at diagnosis into four categories; 0–44 (the reference), 45–59, 60–74, and 75+ years.
- It is estimated that individuals aged 75+ at diagnosis experience 2.25 times higher risk of death due to colon carcinoma than individuals aged 0–44 at diagnosis, a difference which is statistically significant ($P < 0.0005$).
- Similarly, individuals aged 60–74 at diagnosis have an estimated 34% higher risk of death due to colon carcinoma than individuals aged 0–44 at diagnosis, a difference which is statistically significant ($P < 0.02$).

Example: Localised colon carcinoma 1975–1994 V

- These significance tests test the pairwise differences and tell us little about the overall association between age and survival – we need to perform a general test.

R code and output

```
> library(car)
> linearHypothesis(fit, c("agegrp45-59","agegrp60-74","agegrp75+"))

Linear hypothesis test

Hypothesis:
agegrp45 - 59 = 0
agegrp60 - 74 = 0
agegrp75  +  = 0

Model 1: restricted model
Model 2: Surv(surv_mm/12, status == "Dead: cancer") ~ sex + agegrp + year8594

  Res.Df Df  Chisq Pr(>Chisq)
1   6272
2   6269  3 175.88  < 2.2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Example: Localised colon carcinoma 1975–1994 VI

- This is a Wald test of the null hypothesis that all age parameters are equal to zero, i.e. that age is not associated with the outcome.
- We see that there is strong evidence against the null hypothesis, i.e. we conclude that age is significantly associated with survival time.
- The Wald test is an approximation to the likelihood ratio test, which compares the likelihood between models.
- To perform a likelihood ratio test we fit a reduced model without age.

Example: Localised colon carcinoma 1975–1994 VII

R code and output

```
> fit2 <- coxph(Surv(surv_mm/12, status=="Dead: cancer") ~ sex + year8594,
                subset=(stage=="Localised"), data=colon)
> anova(fit,fit2,test="Chisq")

Analysis of Deviance Table
 Cox model: response is  Surv(surv_mm/12, status == "Dead: cancer")
 Model 1: ~ sex + agegrp + year8594
 Model 2: ~ sex + year8594
  loglik  Chisq Df P(>|Chi|)
1 -14342
2 -14430 176.71  3 < 2.2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

- The log likelihood for the model containing age is -14342; for the model excluding age it is -14430.
- The likelihood ratio test statistic for the association of age with survival is calculated as $2 \times (-14342 - (-14430)) = 177$, which is compared to a χ^2 distribution with 3 degrees of freedom (P=0.0001).
- We see that the Wald test statistic (175.88) is very similar in value to the likelihood ratio test statistic (176.71).

We might choose to model age as a continuous variable I

R code and output

```
> fit <- coxph(Surv(surv_mm/12, status=="Dead: cancer") ~ sex + age + year8594,
               subset=(stage=="Localised"), data=colon)
> summary(fit)

  n= 6274, number of events= 1734

                             coef exp(coef)  se(coef)      z Pr(>|z|)
sexFemale                -0.102884  0.902232  0.049362 -2.084   0.0371 *
age                       0.033624  1.034196  0.002342 14.359  < 2e-16 ***
year8594Diagnosed 85-94 -0.290566  0.747840  0.049343 -5.889 3.89e-09 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

                         exp(coef) exp(-coef) lower .95 upper .95
sexFemale                   0.9022     1.1084    0.8190    0.9939
age                         1.0342     0.9669    1.0295    1.0390
year8594Diagnosed 85-94     0.7478     1.3372    0.6789    0.8238
```

We might choose to model age as a continuous variable II

- For each (and every) one year increase in age at diagnosis, we estimate that mortality is 3.4% higher.
- For a 10-year increase in age at diagnosis the estimated hazard ratio is $1.034^{10} = 1.40$.

Comparison of Cox regression to Poisson regression for the analysis of cohort studies I

- The methods are very similar; the basic formulation of both models is

$$\log(\text{rate}) = \boldsymbol{X}\beta.$$

- In both cases, the β parameters are interpreted as log rate ratios.
- Both models are multiplicative (i.e. both assume proportional hazards).
- That is, if the RR for males/females is 3 and the RR for smokers to non-smokers is 4, then the RR for male smokers to female non-smokers is 12 (in a model with no interaction terms).
- In Poisson regression, follow-up time is classified into bands and a separate rate parameter is estimated for each band (or smoothed!), thereby allowing for the possibility that the rate is changing with time.
- It is assumed that the rate is constant within each band, so if the rate is changing rapidly with time we may have to choose very narrow bands.
- In Cox regression, we essentially choose bands of infinitesimal width; each band is so narrow that it includes only a single event.

Comparison of Cox regression to Poisson regression for the analysis of cohort studies II

- Unlike in Poisson regression, we do not estimate the baseline rates within each time band; instead, we estimate the relative rates for the different levels of the covariates.
- As such, if estimating the effect of time is of interest then Poisson regression is a more natural choice.
- Time-by-covariate interactions (i.e., non-proportional hazards) and multiple time scales are, in practice, easier to model in the framework of Poisson regression.

Equivalence of Cox and Poisson regression I

- The Cox model can be viewed as extending the life-table approach *ad absurdum* by:
  1. splitting time as finely as possible,
  2. modelling one covariate, the time-scale, with one parameter per observed value of time,
  3. profiling these parameters out by maximizing the profile likelihood
- Subsequently recover the effect of the timescale by smoothing an estimate of the parameters that was profiled out!

Equivalence of Cox and Poisson regression II

R code

```
library(xtable)
cuts <- (1:20)*12 ## yearly split
localised <- survSplit(Surv(surv_mm, status=="Dead: cancer") ~ sex + agegrp + year8594,
             subset=(stage=="Localised"), data=colon, cut=cuts, episode="timeband") %>%
    group_by(sex,agegrp,year8594,timeband) %>%
    summarise(event=sum(event), pt = sum(surv_mm - tstart)) # collapse
fit <- glm(event ~ sex + agegrp + year8594 + factor(timeband) + offset(log(pt)),
           data=localised,
           family=poisson, control=list(maxit = 200, epsilon=1e-12))
cuts <- with(subset(colon, stage=="Localised" & status=="Dead: cancer"),
             sort(unique(surv_mm))) # split by event times
cuts <- cuts[-length(cuts)]
localised <- survSplit(Surv(surv_mm, status=="Dead: cancer") ~ sex + agegrp + year8594,
                       subset=(stage=="Localised"), data=colon, cut=cuts,
                       episode="timeband") %>%
    group_by(sex,agegrp,year8594,timeband) %>%
    summarise(event=sum(event), pt = sum(surv_mm - tstart)) # collapse
fit2 <- glm(event ~ sex + agegrp + year8594 + factor(timeband) + offset(log(pt)),
           data=localised,
           family=poisson, control=list(maxit = 200, epsilon=1e-12))
fit3 <- coxph(Surv(surv_mm, status=="Dead: cancer") ~ sex + agegrp + year8594,
              subset=(stage=="Localised"), data=colon)
xtable(data.frame(poisson=coef(fit)[2:6], poisson.fine=coef(fit2)[2:6], coxph=coef(fit3)))
```

Equivalence of Cox and Poisson regression III

	sexFemale	agegrp45-59	agegrp60-74	agegrp75+	year8594Diagnosed 85-94
poisson	-0.0930	-0.0503	0.2959	0.8280	-0.2789
poisson.fine	-0.0889	-0.0525	0.2904	0.8093	-0.2814
coxph	-0.0894	-0.0520	0.2924	0.8141	-0.2825

Choice of time scale in a Cox model I

- In Cox regression the hazard ratio (HR) compares the hazard rates of two groups at each t.
- So for any t, the hazards are assumed to be proportional by a multiple, HR.
- Since the comparison is made at each t, the HR is automatically adjusted for t, the underlying time scale.
- For example, if we use attained age as the underlying time scale, then all the HRs will be adjusted for age in the Cox model.
- If age is the chosen as the underlying time scale in the Cox model, the effect of age cannot be estimated directly, since it is incorporated in the shape of the baseline hazard, which is allowed to vary freely.
- The adjustment of the underlying time scale in a Cox model is very efficient, since it adjusts for time in very small intervals.
- Hence, if one of the three possible time scales for your data (time-in-study, attained age, calendar time) is a strong confounder of your exposure–outcome association, then that time scale should be preferred.

Choice of time scale in a Cox model II

- For example, consider a cohort study where individuals are randomly selected from the population and followed for cancer incidence.
- For most cancers, age is a strong confounder. If we are not interested in estimating the effect of age, an efficient approach to adjust for age is to choose age as the underlying time scale [8, 1, 2].
- Thiébaut and Bénichou [8] recommend using age as the timescale and conclude 'we strongly recommend not using time-on-study as the time-scale for analysing epidemiologic cohort data [where entry has no clinical or biological relevance]'.

Analysing the diet data using Cox regression I

- Use attained age as the timescale.

R code and output

```
> scale <- 365.24
> fit <- coxph(Surv((doe-dob)/scale, (dox-dob)/scale, chd) ~ hieng,
             data=diet)
> summary(fit)

             coef exp(coef) se(coef)      z Pr(>|z|)
hienghigh -0.6114    0.5426   0.3028 -2.019   0.0435 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

          exp(coef) exp(-coef) lower .95 upper .95
hienghigh    0.5426      1.843    0.2997    0.9823

Concordance= 0.584  (se = 0.038 )
Rsquare= 0.012   (max possible= 0.755 )
Likelihood ratio test= 4.2  on 1 df,   p=0.04047
Wald test            = 4.08  on 1 df,   p=0.04348
Score (logrank) test = 4.2  on 1 df,   p=0.04033
```

Analysing the diet data using Cox regression II

- This is a test of equality of CHD mortality rates between individuals with a high and low energy intake, adjusted for attained age, and assuming proportional hazards with respect to attained age.
- That is, it is a very similar test to that performed in the framework of Poisson regression (and to the log-rank test).
- The effect estimate and P-value for the test of the effect of `hieng` are very similar in the Cox and Poisson regression models.
- A slight difference is that attained age was categorised in the Poisson regression model and the rate assumed to be constant within each category.

Summary so far

- We have introduced the Cox model to model survival data.
- The Cox model is an alternative to the Poisson regression model.
- The Cox model does not assume a shape of the baseline hazard, but allows it to vary freely.
- The Cox model assumes proportional hazards.
- We need to assess the appropriateness of the proportional hazards assumption.

Assessing the appropriateness of the proportional hazards assumption I

- The proportional hazards assumption is a strong assumption and its appropriateness should generally be assessed.
- The model assumes that the *ratio* of the hazard functions for any two patient subgroups (i.e. two groups with different values of the explanatory variable $\boldsymbol{X}$) is constant over follow-up time.
- Note that it is the hazard ratio which is assumed to be constant. The hazard can vary freely with time.
- When comparing an aggressive therapy vs a conservative therapy, for example, it is not unusual that the patients receiving the aggressive therapy do worse earlier, but then have a lower hazard (i.e. better survival) than those receiving the conservative therapy.
- In this situation, the ratio of the hazard functions will not be constant over time, as is assumed by the PH model.

Assessing the appropriateness of the proportional hazards assumption II

- On Day 1, we showed an example of non-proportional hazards, although this may not be obvious to the untrained eye; it is difficult to assess the PH assumption by looking at the estimates of the survivor function.
- If the hazard functions cross, it is possible that the effect of treatment will not be statistically significant despite the presence of a clinically interesting effect.
- As such, it is important to plot survival curves before fitting the model and to assess the appropriateness of the proportional hazards assumption of the proportional hazards assumption after the model has been fitted.
- Note that the hazard functions do not have to cross for the PH assumption to be violated. For example, a hazard ratio of 4 which gradually decreases with time to a value of 1.5 is an example of non-proportional hazards.
- Hess (1995) [6] reviews methods for assessing the appropriateness of the proportional hazards assumption.
- Therneau & Grambsch [7] give a more up-to-date review and include code for implementing the various methods in SAS and R.

Assessing the appropriateness of the proportional hazards assumption III

- Following is a list of commonly used methods for assessing the appropriateness of the proportional hazards assumption (in increasing order of utility):
  1. Plotting the cumulative survivor functions and checking that they do not cross. This method is not recommended, since the survivor functions do not have to cross for the hazards to be non-proportional.
  2. Plotting the log cumulative hazard functions over time and checking for parallelism. This method does not provide any formal fit criteria and is more descriptive.
  3. Including time-by-covariate interaction terms in the model and testing statistical significance. For example, a statistically significant time-by-exposure term would indicate a trend in the hazard ratio with time.
  4. Plotting Schoenfeld's residuals against time to identify patterns.
- The first two methods do not allow for the effect of other covariates, whereas the second two methods do.
- Including a time-by-covariate interaction in the model has the advantage that we obtain an estimate of the hazard ratio as a function of time.

Modelling interactions with time to test and model non-proportional hazards I

- Non proportional hazards is just a special name for 'effect modification by time on a log scale'.
- Effect modification is a familiar concept; we can use interaction terms to test for effect modification and to estimate the effect of exposure in each stratum of the modifier.
- Note that Poisson regression also assumes proportional hazards. To allow for non-proportional hazards we fit time by covariate interaction effects.
- The difficulty with the Cox model is that we don't explicitly estimate the effect of time so it's not obvious how to fit a time by covariate interaction.

Modelling interactions with time to test and model non-proportional hazards II

- We can use one of two approaches:
  - Split by time.
  - Use the options in R for modelling 'time-varying covariates' (using the `tt()` function in `coxph()`).
- What we are actually interested in is the situation where the *effect* of a covariate varies by time, which is not the same as the value of covariate varying with time. We'll discuss the distinction in more detail on slide 72.
- We do not explicitly estimate the the effect of the underlying time scale in a Cox model, but we can estimate interactions with the underlying time scale.
- Note that it is possible to estimate the underlying time-scale (baseline cumulative hazard and hazard) after fitting a Cox model (see `survival::basehaz` and `biostat3::coxphHaz`).
- We still allow the baseline hazard to vary freely, but relax the assumption that hazards must be proportional over time.

Modelling interactions with time, by splitting time

- One way to model interactions with the underlying time scale, is to split time and allow covariates to have different effects over time.
- The R function `survSplit()` divides risktime into several records, one for each timeband we specify.
- We will now model an interaction with time in the colon carcinoma data, to allow for different hazard ratios for calendar period before and after 2 years (24 months).

Colon data: estimating a time by period interaction I

- We have seen that mortality depends on calendar period of diagnosis (HR 0.72 for recent/early period).
- Would we expect mortality in the recent period to be 28% lower at all points in the follow-up or is it conceivable that the effect is greater (or even restricted) to the period immediately following diagnosis?
- If the effect is different early in the follow-up, compared to later in the follow-up, then we have a case of non-proportional hazards.
- That is, the effect of calendar period is modified by time since diagnosis.
- Based on clinical knowledge, we choose to estimate the effect separately for the first 24 months of follow-up.

Colon data: estimating a time by period interaction II

- We start with splitting the data on time, $t < 24$ months, using `survival::survSplit`.

R code

```
> localised <- survSplit(Surv(surv_mm, status=="Dead: cancer")~ agegrp+sex+year8594,
                         cut=c(24,1000),
                         data=colon, subset=(stage=="Localised"),
                         episode="timeband")
> localised <- transform(localised, timeband = factor(timeband))
```

- We can now fit a model containing the interaction between year of diagnosis (two categories) and time (in two categories).

Colon data: estimating a time by period interaction III

R code and output

```
> summary(coxph(Surv(tstart,surv_mm,event)~agegrp+sex+year8594*timeband,
                data=localised))

  n= 10885, number of events= 1734

                                    coef exp(coef) se(coef)      z Pr(>|z|)
agegrp45-59                      -0.05169   0.94962  0.13845 -0.373  0.70888
agegrp60-74                       0.29122   1.33806  0.12573  2.316  0.02055 *
agegrp75+                         0.81496   2.25908  0.12605  6.465 1.01e-10 ***
sexFemale                        -0.09003   0.91390  0.04937 -1.824  0.06822 .
year8594Diagnosed 85-94          -0.42272   0.65526  0.06531 -6.473 9.63e-11 ***
timeband2                              NA        NA  0.00000     NA       NA
year8594Diagnosed 85-94:timeband2  0.32288   1.38110  0.09883  3.267  0.00109 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Warning message:
In coxph(Surv(tstart, surv_mm, event) ~ agegrp + sex + year8594 *  :
  X matrix deemed to be singular; variable 6
```

Colon data: estimating a time by period interaction IV

- Recall how we interpret interaction effects (in general).
  - `year8594Diagnosed 85-94`; effect of the later calendar period of diagnosis (1985–1994)
  - `timeband2`; effect of time in the second period of follow-up (after 24 months).
  - `year8594Diagnosed 85-94:timeband2`; additional (multiplicative) effect of later calendar period (1985–1994) at the second period of follow-up (after 24 months).

Colon data: estimating a time by period interaction V

- `timeband2` does not have the usual interpretation because we have already adjusted for the effect of time since diagnosis (as the underlying timescale).
- We are effectively trying to adjust for the same confounder in two different ways in the same model. We should ignore this estimate and focus on the other two.
- The estimated hazard ratio for the effect of period of diagnosis is
  - 0.72 when assuming proportional hazards
  - 0.66 for the early period
  - 0.91 for the recent period ($0.655 \times 1.381 = 0.90$)
- We see that there is evidence that the effect of period of diagnosis is more pronounced early in the follow-up.
- If the interaction effect was zero (HR associated with `year8594Diagnosed 85-94:timeband2` equal to one) then there would be no effect modification (proportional hazards).
- We can see that the interaction effect is statistically significant (p=0.001).
- As we saw previously, we can reparameterise the model to estimate the effect of period within each timeband.

Colon data: estimating a time by period interaction VI

- Then we fit the model again using these indicator variables for the effect of calendar period over time.

R code and output

```
> localised <- transform(localised,
                         later=ifelse(year8594=="Diagnosed 85-94",1,0))
> summary(coxph(Surv(tstart,surv_mm,event)~agegrp+sex+later:timeband,
                data=localised))

                    coef exp(coef) se(coef)      z Pr(>|z|)
agegrp45-59     -0.05169   0.94962  0.13845 -0.373   0.7089
agegrp60-74      0.29122   1.33806  0.12573  2.316   0.0205 *
agegrp75+        0.81496   2.25908  0.12605  6.465 1.01e-10 ***
sexFemale       -0.09003   0.91390  0.04937 -1.824   0.0682 .
later:timeband1 -0.42272   0.65526  0.06531 -6.473 9.63e-11 ***
later:timeband2 -0.09984   0.90498  0.07438 -1.342   0.1795
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

- Note that we could also use the `biostat3::lincom()` function rather than reparameterising the model.

Colon data: estimating a time by period interaction VII

- The estimated hazard ratio, based on the above model, for patients diagnosed 1985–94 compared to 1975–84 is 0.655 for the period up to 2 years of follow-up and 0.905 for the period after 2 years of follow-up (as we previously saw).
- To test if this interaction is statistically significant we could perform a LR test, comparing the model with the interaction to the model without the interaction.

Colon data: estimating a time by period interaction VIII

R code and output

```
> fit <- coxph(Surv(tstart,surv_mm,event)~agegrp+sex+
                   year8594*timeband,
               data=localised)
> anova(fit,test="Chisq")
Terms added sequentially (first to last)

                   loglik   Chisq Df Pr(>|Chi|)
NULL               -14442
agegrp             -14360 163.661  3  < 2.2e-16 ***
sex                -14358   2.784  1   0.095209 .
year8594           -14342  32.681  1  1.086e-08 ***
timeband           -14342   0.000  0   1.000000
year8594:timeband -14337  10.648  1   0.001102 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Colon data: estimating a time by period interaction IX

- Note that the previous z test statistic (slide 63) was 3.27. If we square this we get a test statistic that is χ^2_1.

$$3.27^2 = 10.69$$

- Both of these tests are testing the hypothesis that the interaction effect is zero versus it is non-zero. The reason for the small difference in the test statistic is that one is a likelihood ratio test and one is a Wald test.

A look at the interaction models (for completeness) I

- Consider again a proportional hazards model with one single binary variable, X_1, which takes the value 1 if an exposure is present and 0 if it is absent

$$\lambda(t|\boldsymbol{X}) = \lambda_0(t)\exp(\beta_1 X_1).$$

- The hazard ratio for exposed to unexposed is given by $\exp(\beta_1)$.
- We now construct a second variable, $X_2 = X_1 t$ and include this in the model, in addition to X_1. The variable X_2 takes the value t if the exposure is present and 0 if it is absent

$$\lambda(t|\boldsymbol{X}) = \lambda_0(t)\exp(\beta_1 X_1 + \beta_2 X_1 t).$$

- Based on this model, the hazard ratio for exposed to unexposed is given by $\exp(\beta_1 + \beta_2 t)$.
- An estimate for β_2 significantly different from 0 indicates that the hazard ratio is non-constant over time. $\beta_2 > 0$ indicates that the hazard ratio increases with time and $\beta_2 < 0$ indicates it decreases with time.

A look at the interaction models (for completeness) II

- This is not a general test of the proportional hazards assumption. It tests against the alternative that the hazard ratio changes monotonically with time.
- Another alternative might be that the hazard ratio is constant for an initial time period, say $t = 2$ years, but takes on a different (constant) value for the remainder of follow-up [5].
- To test against this alternative, we construct a variable X_2 which takes the value 1 if the exposure is present and $t > 2$ years, and 0 otherwise.
- In the resulting model containing the variables X_1 and X_2, the hazard ratio for exposed to unexposed for the period $t \leq 1$ year is given by $\exp(\beta_1)$ and for $t > 2$ years it is given by $\exp(\beta_1 + \beta_2)$.
- An estimate for β_2 significantly different from 0 indicates that the hazard ratio is different between the two time periods.

Time-varying covariates I

- We have been considering the situation where the *effect* of a covariate varies with time.
- It is possible that the underlying *values* of covariates can change during follow-up. For example, blood pressure, occupational exposure to carcinogens, parity, CD4 count, or cumulative exposure to cigarettes.
- Another application is in observational studies where an intervention may occur at any point in the follow-up. At the time of the intervention, the explanatory variable associated with the intervention changes value from 0 (false) to 1 (true).
- We highly recommend the time-splitting approach for modelling such data.
- That is, we split to obtain a separate observation at every value of the time-varying covariate.
- Exercise 22 examines a possible effect of *marital bereavement* (loss of husband or wife) on all–cause mortality in the elderly (see Clayton & Hills, §32.2).

Time-varying covariates II

- Bereavement is a time-varying exposure – all subjects enter as not bereaved but may become bereaved at some point during follow–up.
- A distinction is made between internal variables (which relate to an individual and can only be measured while a patient is alive) and external variables (which do not necessarily require survival of the patient for their existence).
- Care should be taken when modelling time-dependent covariates, particularly with internal variables [4, 9].

The `tt` option in `coxph` I

- The `tt()` functions in `coxph` can also be used for estimating time-varying effects of covariates.
- Let's again fit the model where we allow the effect of period to differ in the first 2 years of follow-up.

The tt option in coxph II

R code and output

```
> colon2 <- transform(colon, later=ifelse(year8594=="Diagnosed 85-94",1,0))
> summary(fit <- coxph(Surv(surv_mm,status=="Dead: cancer")~agegrp+sex+year8594+tt(later),
                data=colon2, subset=(stage=="Localised"),
                tt = function(x, t, ...) x*(t>=24)))

                         exp(coef) exp(-coef) lower .95 upper .95
agegrp45-59                 0.9496     1.0531    0.7239    1.2457
agegrp60-74                 1.3381     0.7474    1.0458    1.7120
agegrp75+                   2.2591     0.4427    1.7646    2.8922
sexFemale                   0.9139     1.0942    0.8296    1.0068
year8594Diagnosed 85-94     0.6553     1.5261    0.5765    0.7447
tt(later)                   1.3811     0.7241    1.1379    1.6763

> lincom(fit, "year8594Diagnosed 85-94+tt(later)", eform=TRUE)

                                   Estimate     2.5 %  97.5 %    Chisq
year8594Diagnosed 85-94+tt(later) 0.9049837 0.7822161 1.04702 1.801605
                                  Pr(>Chisq)
year8594Diagnosed 85-94+tt(later)  0.1795186
```

The tt option in coxph III

- The cutoff at 24 months was chosen arbitrarily. For the first two years of follow-up the estimated hazard ratio was 0.656, while after two years the hazard ratio was $0.656 \times 1.381 = 0.905$.
- Choosing the cutpoint after inspection of the data will invalidate statistical inference (i.e. reported P-values will be too low).
- We have examined only one possible alternative to proportional hazards (a step function with a single step at 24 months).
- In practice, it is possible to fit any model of the form

$$\lambda(t|\boldsymbol{X}) = \lambda_0(t)\exp(\beta_1 X_1 + \beta_2 X_1 f(t)),$$

where $f(t)$ is a function of time.

Tests of the PH assumption based on Schoenfeld residuals

- If the PH assumption holds then the Schoenfeld residuals (a diagnostic specific to the Cox model) should be independent of time.
- In its simplest form, when there are no ties, the unscaled Schoenfeld residual for covariate $x_u, u = 1, ..., p$, and for observation j observed to fail is

$$r_{uj} = x_{uj} - \frac{\sum_{i \in R_j} x_{ui} \exp(\mathbf{x}_i \hat{\beta}_{\mathbf{x}})}{\sum_{i \in R_j} \exp(\mathbf{x}_i \hat{\beta}_{\mathbf{x}})}$$

- That is, r_{uj} is the difference between the covariate value for the failed observation and the weighted average of the covariate values over all those subjects at risk of failure when subject j failed.
- A test of the PH assumption can be made by modelling the scaled Schoenfeld residuals as a function of time and testing the hypothesis of a zero slope.

Application to localised colon carcinoma I

R code and output

```
fit <- coxph(Surv(surv_mm/12,status=="Dead: cancer")~sex+agegrp+year8594,
+              data=colon, subset=(stage=="Localised"))
> ## redefine the print method (hack)
> print.cox.zph <- function(object, ...)
+     printCoefmat(object$table, ...)
> cox.zph(fit) # survival transformation
                             rho      chisq      p
sexFemale                0.0035758  0.0222461 0.8814
agegrp45-59              0.0101412  0.1782300 0.6729
agegrp60-74              0.0087225  0.1328470 0.7155
agegrp75+               -0.0486362  4.1019081 0.0428
year8594Diagnosed 85-94  0.0826805 12.1550765 0.0005
GLOBAL                        NA   47.3058019 0.0000
> rm(print.cox.zph) # tidy up
```

- The tests suggest that there is evidence that the hazards are nonproportional by calendar period (and possibly age).

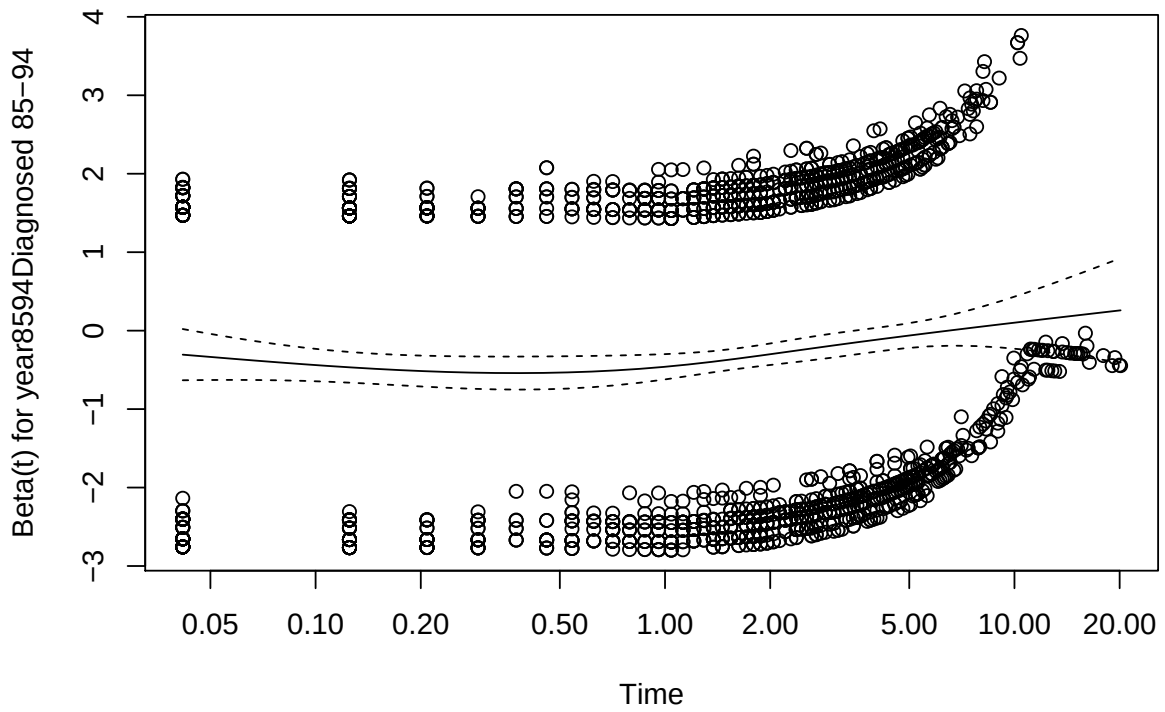
Application to localised colon carcinoma II

- Rather than just fitting a straight line to the residuals and testing the hypothesis of zero slope (as is done by `cox.zph`) we can study a plot of the residuals along with a smoother to assist us in determining how the mean residual varies as a function of time.
- The smooth illustrates how the log hazard ratio varies as a function of time. We see, for example, that the effect of period is larger during the initial years of follow-up.

Application to localised colon carcinoma III

R code and output

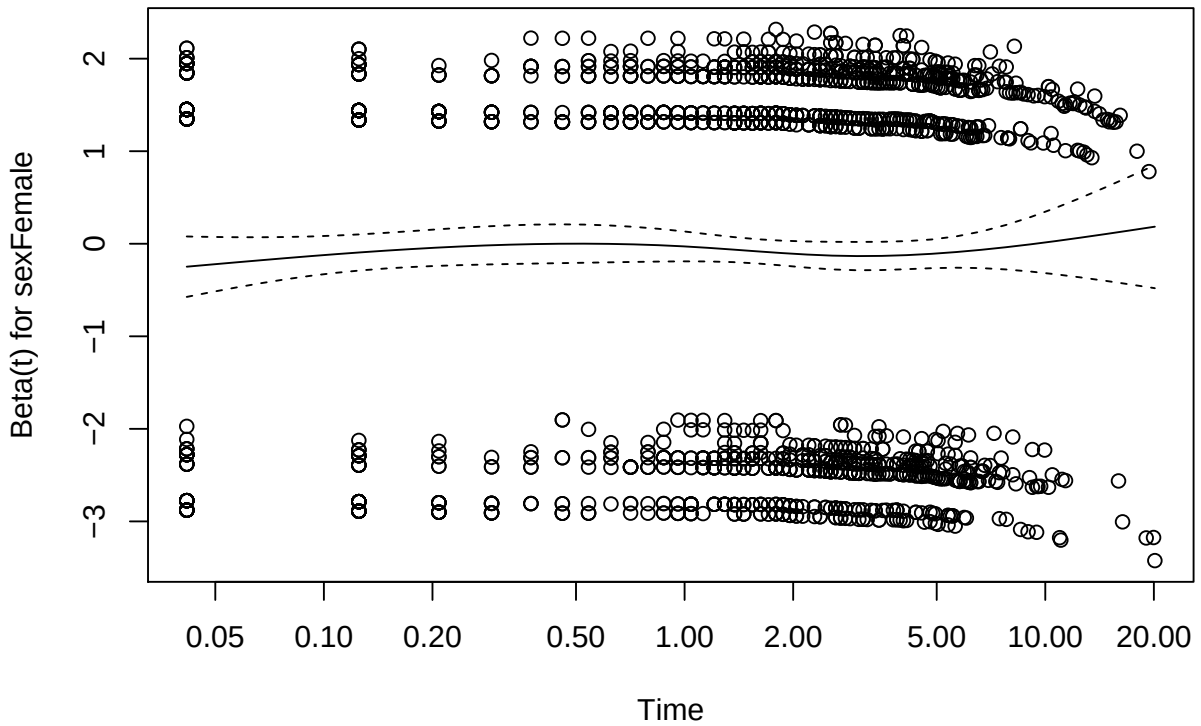
```
> fit <- coxph(Surv(surv_mm,status=="Dead: cancer")~sex+agegrp+year8594,
               data=colon, subset=(stage=="Localised"))
> plot(cox.zph(fit,transform=log)[5])
```



Application to localised colon carcinoma IV

R code and output

```
> plot(cox.zph(fit,transform=log)[1])
```



A model including stage I

R code and output

```
> known <- transform(colon, stage=droplevels(stage, "Unknown"))
> fit <- coxph(Surv(surv_mm/12,status=="Dead: cancer")~
                  sex+agegrp+stage+year8594,
               data=known)
> summary(fit)

  n= 13208, number of events= 7186
   (2356 observations deleted due to missingness)

                            coef exp(coef) se(coef)      z Pr(>|z|)
sexFemale                -0.04651   0.95456  0.02437 -1.908   0.0564 .
agegrp45-59               0.08546   1.08922  0.06382  1.339   0.1806
agegrp60-74               0.27355   1.31462  0.05868  4.662 3.13e-06 ***
agegrp75+                 0.62357   1.86557  0.05937 10.504  < 2e-16 ***
stageRegional             0.83689   2.30916  0.04109 20.367  < 2e-16 ***
stageDistant              2.11896   8.32251  0.02937 72.150  < 2e-16 ***
year8594Diagnosed 85-94 -0.15366   0.85756  0.02399 -6.406 1.49e-10 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

- Stage is categorised into Localised, Regional and Distant tumours.

A model including stage II

R code and output

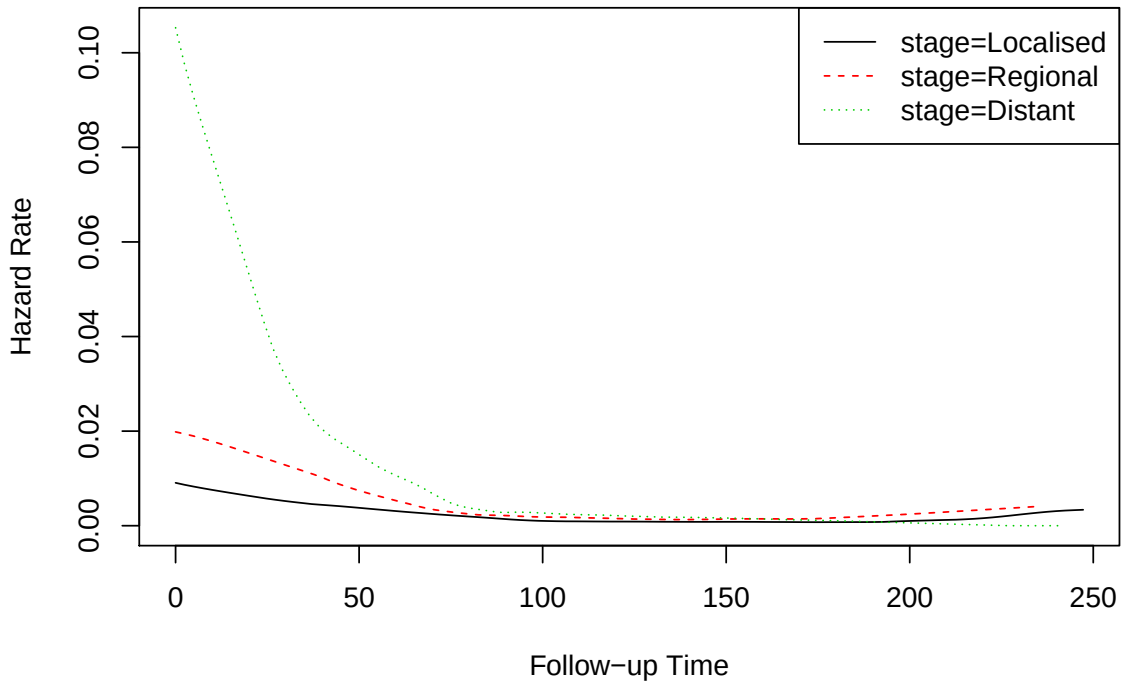
```
> cox.zph(fit,transform=log)
                              rho       chisq      p
sexFemale                0.0095729   0.6612137 0.4161
agegrp45-59             -0.0070487   0.3572841 0.5500
agegrp60-74             -0.0103022   0.7643971 0.3820
agegrp75+               -0.0572602  23.4539894 0.0000
stageRegional            0.0211593   3.1912659 0.0740
stageDistant            -0.0825732  44.9546333 0.0000
year8594Diagnosed 85-94  0.0094099   0.6431899 0.4226
GLOBAL                          NA 173.5798931 0.0000
```

- There is evidence that the hazards are heavily non-proportional by stage.
- A plot of the empirical hazards (slide 84) suggests that individuals diagnosed with distant metastases have proportionally much higher mortality early in the follow-up but once they have survived several years their mortality is not that much higher than the other age groups.
- The plots of the fitted hazards (slide 85) show the effect of the assumption of proportional hazards.

A model including stage III

R code and output

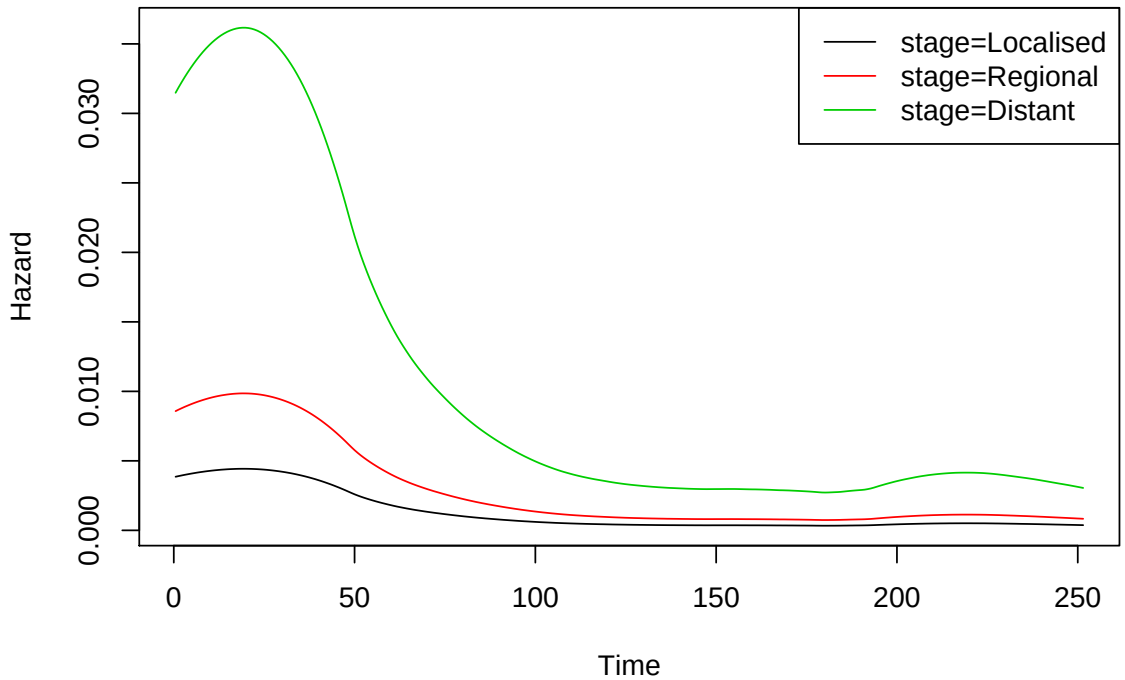
```
> fit <- muhaz2(Surv(surv_mm,status=="Dead: cancer")~stage,
                data=known)
> plot(fit)
```



A model including stage IV

R code and output

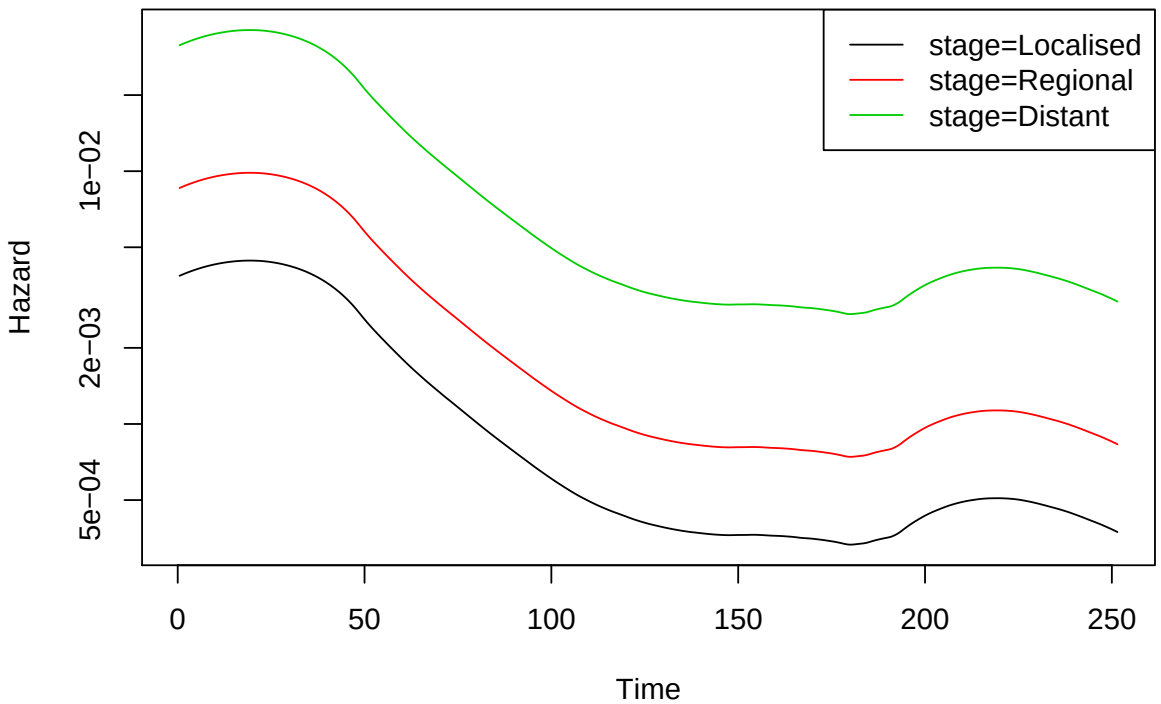
```
> fit <- coxph(Surv(surv_mm,status=="Dead: cancer")~stage,
               data=known)
> plot(coxphHaz(fit,newdata=data.frame(stage=levels(known$stage))))
```



A model including stage V

R code and output

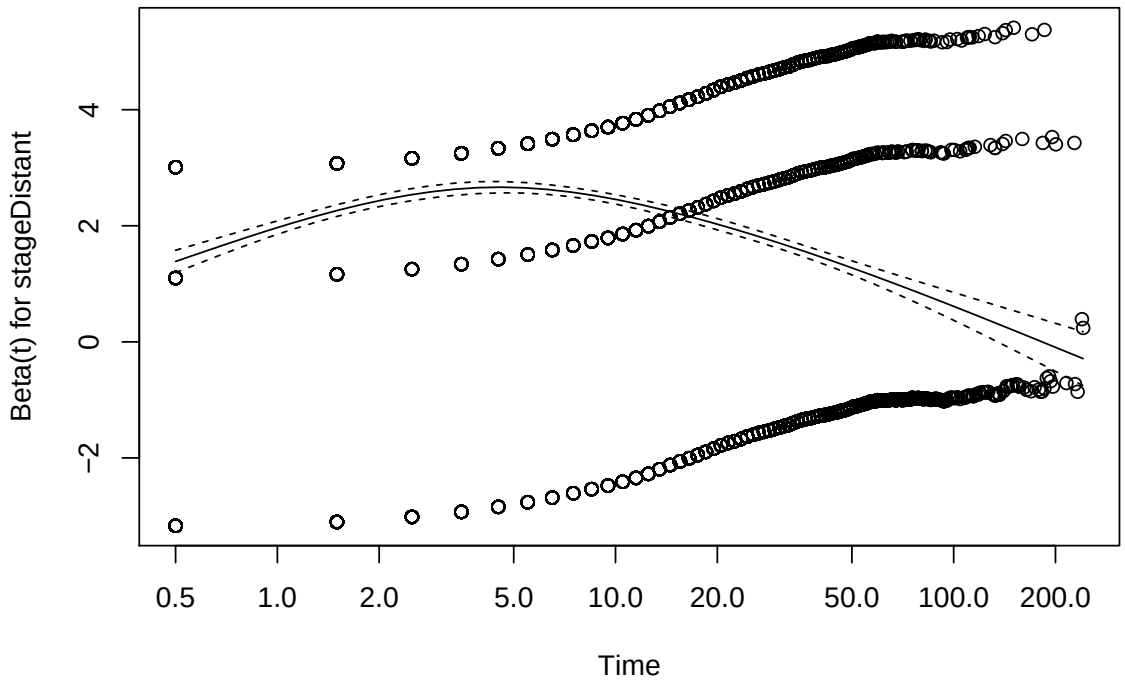
```
> plot(coxphHaz(fit,newdata=data.frame(stage=levels(known$stage))),
       log="y")
```



A model including stage VI

R code and output

```
> fit <- coxph(Surv(surv_mm,status=="Dead: cancer")~stage,
                data=known)
> plot(cox.zph(fit)[2])
```



The stratified Cox model I

- The Cox model assumes that the baseline hazard (e.g., instantaneous mortality rate in the reference group) is an arbitrary function of time.
- The hazard functions for each of the other groups are assumed to be proportional to the baseline.
- It is possible to relax this assumption to allow separate baseline hazards for different groups, say for each level of age at diagnosis.
- This is known as a stratified proportional hazards model and is a useful method for modelling data where non-proportional hazards are suspected for a factor that is not of primary interest.
- A model stratified on agegrp is analogous to including an agegrp:time interaction in a Poisson regression model.
- Use the `strata()` term in the `coxph` formula to specify the strata variables.
- The mathematical formula is:

$$\lambda(t|x, j) = \lambda_j(t) \exp(\beta x)$$

where j is an index for the strata.

The stratified Cox model II

R code and output

```
> fit <- coxph(Surv(surv_mm/12,status=="Dead: cancer")~
                  sex+year8594+strata(agegrp),
               data=colon, subset=(stage=="Localised"))
> summary(fit)

  n= 6274, number of events= 1734

                            coef exp(coef) se(coef)      z Pr(>|z|)
sexFemale                -0.08958   0.91431  0.04938 -1.814   0.0697 .
year8594Diagnosed 85-94 -0.28200   0.75427  0.04942 -5.707 1.15e-08 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Summary of Day 3 I

- We have introduced the Cox proportional hazards regression model and shown how it is very similar to Poisson regression.
- The Cox model assumes proportional hazards (as does Poisson regression), which means that the estimated HRs between groups are constant over time, although we can relax this assumption by modelling interactions.
- The proportional hazards assumption can be tested by fitting time by covariate interactions, which allows effects to vary over time.
- The PH assumption in Cox regression can also be tested using scaled Schoenfeld residuals.

Summary of Day 3 II

- Poisson regression models assume constant hazards or piecewise constant hazards over time, whereas the Cox model allows the hazard to vary freely over time.
- Can make Poisson regression more 'Cox-like' by making the pieces smaller.
- Hazard ratios from a Cox model are automatically adjusted for confounding by the underlying time scale. One should choose an appropriate timescale.

Exercises for Friday afternoon

9. Localised melanoma: modelling cause-specific mortality using Cox regression. [This is a key exercise]
10. Examining the proportional hazards hypothesis (localised melanoma). [This is also a key exercise]
11. Cox regression with observed (all-cause) mortality as the outcome.
12. Cox model for cause-specific mortality for melanoma (all stages).
13. Modelling the diet data using Cox regression.

References I

N. E. Breslow, J. H. Lubin, and B. Langholz.
Multiplicative models and cohort analysis.
Journal of the American Statistical Association, 78, 1983.

Yin Bun Cheung, Fei Gao, and Kei Siong Khoo.
Age at diagnosis and the choice of survival analysis methods in cancer epidemiology.
J Clin Epidemiol, 56(1):38–43, Jan 2003.

D. R. Cox.
Regression models and life tables (with discussion).
Journal of the Royal Statistical Society. Series B, 34:187–220, 1972.

L. D. Fisher and D. Y. Lin.
Time-dependent covariates in the cox proportional-hazards regression model.
Annu Rev Public Health, 20:145–57, 1999.

Miguel A. Hernán.
The hazards of hazard ratios.
Epidemiology, 21(1):13–15, Jan 2010.

Kenneth R. Hess.
Graphical methods for assessing vioations of the proportional hazards assumption in Cox regression.
Statistics in Medicine, 14:1707–1723, 1995.

T. M. Therneau and P. M. Grambsch.
Modelling Survival Data: Extending the Cox Model.
Springer: New York, 2000.

References II

Anne C M Thiébaut and Jacques Bénichou.
Choice of time-scale in Cox's model analysis of epidemiologic cohort data: a simulation study.
Stat Med, 23(24):3803–3820, Dec 2004.

R. A. Wolfe and R. L. Strawderman.
Logical and statistical fallacies in the use of Cox regression models.
Am J Kidney Dis, 27:124–9, 1996.