Introduction to Cox regression

Paul Dickman

September 2020
www.pauldickman.com

Overview of this lecture

Target audience is students and researchers in biomedical sciences without extensive training in statistics.

- The previous lecture – introduction to survival analysis – covered:
  1. The survival function, hazard function, and how they are related.
  2. Estimating the survival function using the Kaplan-Meier method.
  3. Testing for differences in survival using the log rank test.
- This lecture will present an introduction to modelling time-to-event data, with focus on Cox regression.
- When modelling survival data, we model the rate (hazard) so we'll start with an introduction to rates.
- Slides available at `http://www.pauldickman.com/video/cox-regression/`
- Examples use R, but Stata and SAS code available on the same page as the slides.

Rates and person-time

- A *rate* is a measure of change in one quantity per unit of another quantity. In biomedical sciences, rates typically have units 'events per unit time'.
  - Mortality rate: 0.5 deaths per 1,000 person-years
  - Incidence rate: 14 cancers per 100,000 person-years
- Mortality rates and incidence rates are *event rates*.
- The term 'hazard rate' (or 'hazard') is the generic term used in survival analysis to describe the 'event rate'. If, for example, the event of interest is disease incidence then the hazard represents the incidence rate.

$$\text{hazard} = \frac{\text{number of events}}{\text{time at risk}}$$

- Time-at-risk is measured in units of person-years or similar (e.g. person-months).

Rates and person-time (2)

- Person-time is a method of measurement combining persons and time; it is used to aggregate the total population at risk assuming that 10 people at risk for one year is equivalent to 1 person at risk for 10 years.
- If five people are followed for one year, they are followed for 5 person-years.
- If two persons are followed for 2.5 years, they are followed for 5 person-years.

Rates and person-time (3)

- If five people are followed for one year, and one experience a cancer, then the incidence rate is $1/5 = 0.2$ cases per person-year.
- If two persons are followed for 2.5 years, and one experience a cancer, then the incidence rate is $1/5 = 0.2$ cases per person-year.
- Often cancer incidences are reported per 100,000 person-years. For example, an incidence rate of 4 per 100,000 person-years is equivalent to 0.04 per 1,000 person-years and 0.00004 per person-year.

Hazard rates and the hazard function, $\lambda(t)$

- In contrast to the survivor function, which describes the probability of *not* failing before time t, the hazard function focuses on the failure rate at time t among those individuals who are alive at time t. So, the survival function is formally defined for a random time variable T by

$$S(t) = \Pr(T > t) = 1 - F(t). \tag{1}$$

  where $F(t)$ is the failure proportion (aka the cumulative density function).

- The hazard function is defined, for a random time variable T, by

$$\lambda(t) = \lim_{\Delta t \to 0} \frac{\Pr(t \leq T < t + \Delta t \mid T \geq t)}{\Delta t} \tag{2}$$

Hazard rates and the hazard function, $\lambda(t)$ (2)

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

- The hazard function, $\lambda(t)$, is the instantaneous event rate at time t, conditional on survival up to time t.
- From Equation 2, one can see that $\lambda(t)\Delta t$ may be viewed as the 'approximate' probability of an individual who is alive at time t experiencing the event in the next small time interval Δt.
- The units are events per unit time.
- The hazard is a rate, not a probability, so $\lambda(t)$ can take any value between zero and infinity, as opposed to $S(t)$ which is restricted to the interval $[0, 1]$.
- A lower value for $\lambda(t)$ implies a higher value for $S(t)$ and *vice versa*.

Choice of time scale

- There are several time scales along which rates might vary. These differ from one another only in the choice of *time origin*, the point at which time is zero.
- Consider the following questions?
  1. What is the time?
  2. How old are you?
  3. For how long have you lived at your current address?
- What is the time origin (when is time zero) for each?
- In which units did you specify time? Could different units have been used?
- Time progresses in the same manner but, in answering these questions, we have applied a different time origin and used different units.

Common time scales in medical research

Origin	Time scale
Birth	Age
A fixed date	Calendar time
First exposure	Time exposed
Entry into study	Time in study
Diagnosis	Time since diagnosis
Start of treatment	Time on treatment

- In many of the methods used in survival analysis, effects are adjusted for the underlying time scale. Choice of time scale therefore has important implications.
- On many time scales, subjects do not enter follow-up at the time origin, $t = 0$. To deal with these issues, the `Surv` function allows for both the entry and exit times to be specified prior to the event indicator.

A sample of 35 patients diagnosed with colon carcinoma during 1985–94; followed-up until the end of 1995

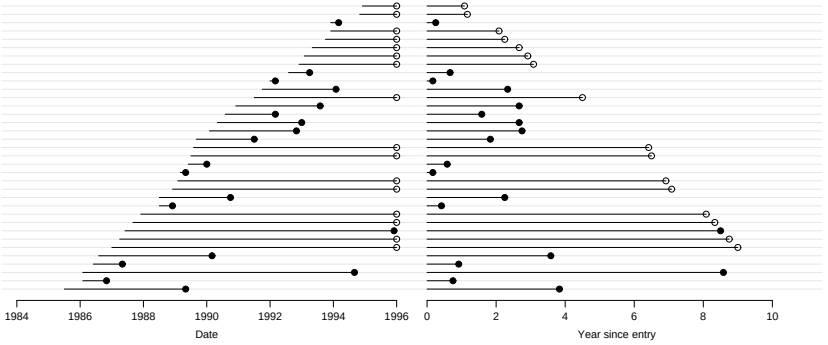
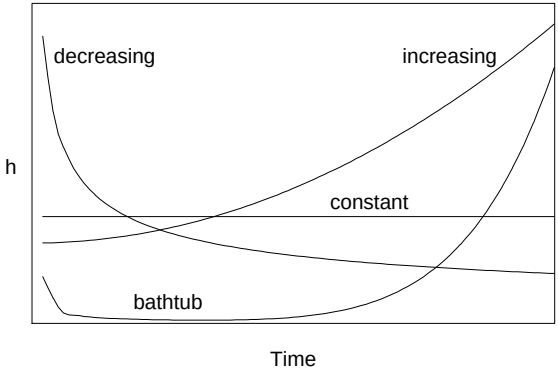


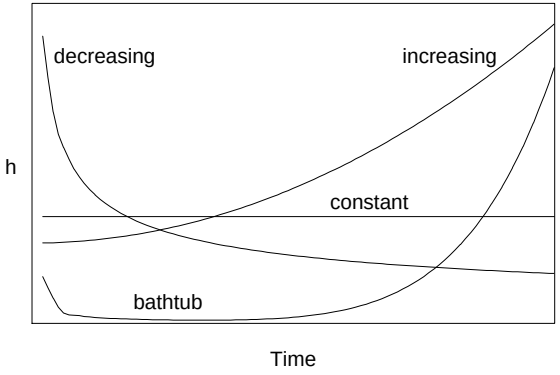
Figure 1: Calendar time (left) and time from entry in years (right)

Common forms for the hazard function



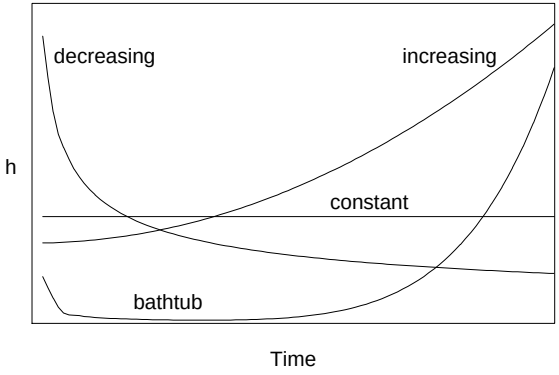
- A bathtub-shaped hazard is appropriate in most human populations followed from birth, where the hazard rate decreases to almost zero after an initial period of infant mortality, and then starts to increase again later in life.
- A decreasing hazard is appropriate following the diagnosis of many types of cancer, where mortality due to the cancer is highest immediately following diagnosis.

Common forms for the hazard function – constant hazards



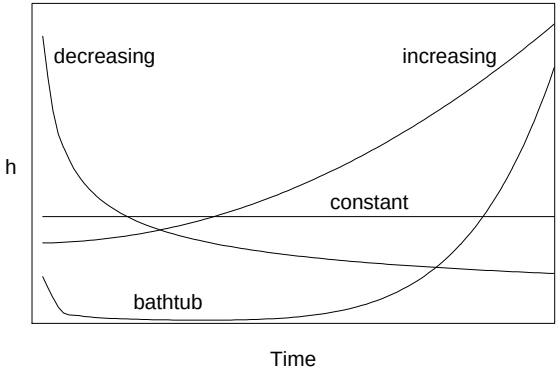
- A constant hazard function is often used for modelling the lifetime of electronic components, but is also appropriate following the diagnosis of some types of cancer, most notably cancers of the breast and prostate, where excess mortality due to the cancer is relatively constant over time.

Common forms for the hazard function – constant hazards



- 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.

Common forms for the hazard function – constant hazards



- An exponential distribution has also been used to model the time between goals in hockey [1].
- The average time to winning a prize for a regular lotto player, for example, can be described by an exponential distribution.

Parametric survival models

- 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.
- The Weibull distribution, which has two parameters, is a more flexible distribution in which the hazard can be either monotonic increasing, decreasing, or constant.
- The Weibull, log-normal and Gompertz distributions have proved to be applicable in several types of medical survival studies, but parametric distributions are often not appropriate.

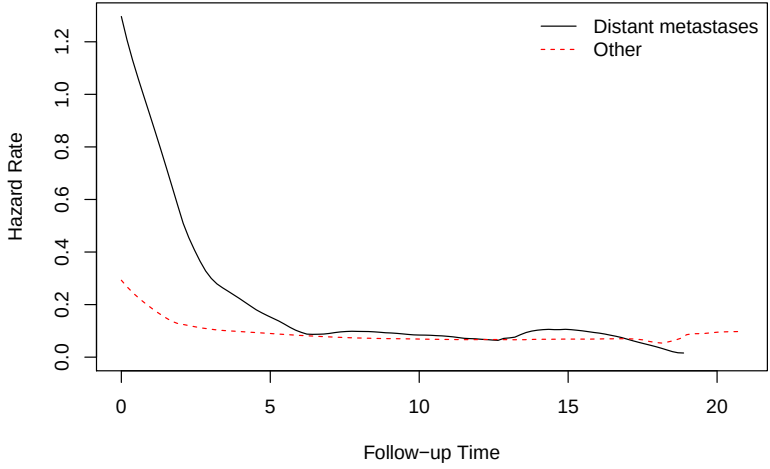
The shape of hazards in a Cox model

- The Cox proportional hazards regression model ('Cox model') does not make any assumption about the shape of the hazard function (Cox 1972 [2]).
- Instead, the baseline hazard is allowed to vary freely.
- The Cox model estimates hazard ratios relative to the baseline hazard.
- The estimated hazard ratios are adjusted for the effect of time, but the baseline hazard is not estimated when fitting the model.

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

- 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: How might we specify a model?



The Cox proportional hazards model

- 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 covariates X is a multiple of the baseline hazard.
- 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 [2], the hazard for individuals with all covariates X at the reference level, is an arbitrary function of time, often called the baseline hazard and denoted by $\lambda_0(t)$.

The Cox proportional hazards model (2)

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

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

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

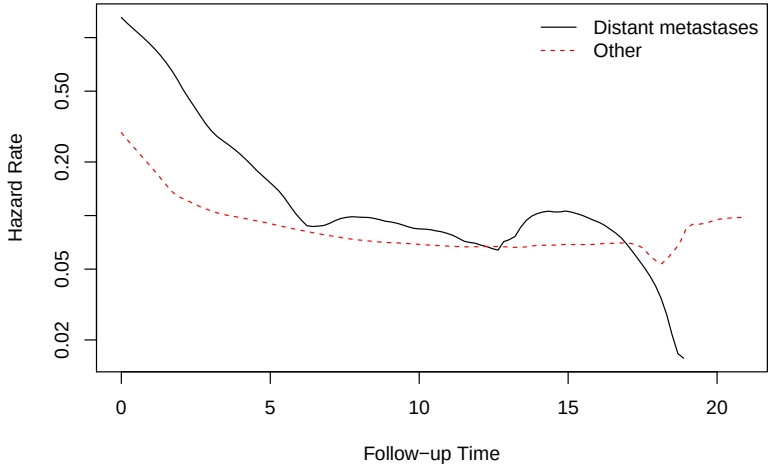
- The difference between two hazards is a constant β regardless of t

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

- That is, the two hazard functions are assumed to be parallel on a log scale.

Empirical hazards on the log scale

- The Cox model assumes that the two hazard curves are parallel on a log scale. Does that assumption appear reasonable?

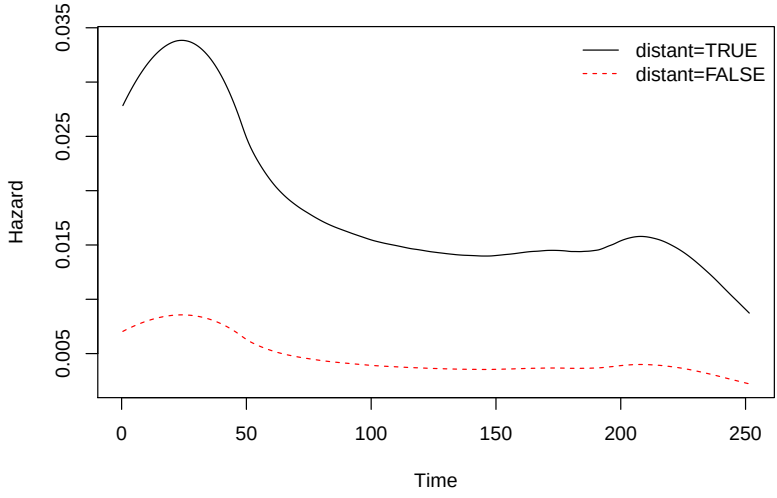


Fit the Cox model in R

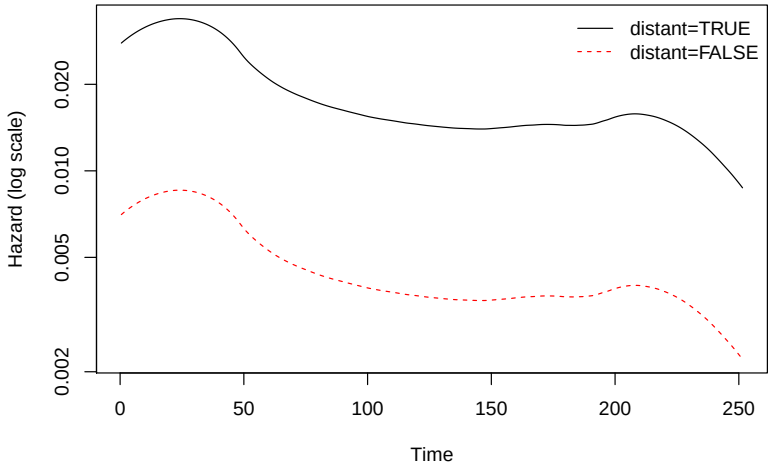
```
> colon2 <- transform(biostat3::colon, distant = (stage == "Distant"),
+                       dead = status %in% c("Dead: cancer"))
>
> summary(coxph(Surv(surv_mm,dead)~distant, data=colon2))
Call:
coxph(formula = Surv(surv_mm, dead) ~ distant, data = colon2)

  n= 15564, number of events= 8369

               coef exp(coef) se(coef)     z Pr(>|z|)
distantTRUE 1.66395   5.28011  0.02292 72.61   <2e-16 ***

            exp(coef) exp(-coef) lower .95 upper .95
distantTRUE      5.28     0.1894     5.048     5.523
```

Fitted hazards

- Although the Cox model does not estimate the hazard functions in the estimation process, we can estimate the predicted hazards after fitting the model.



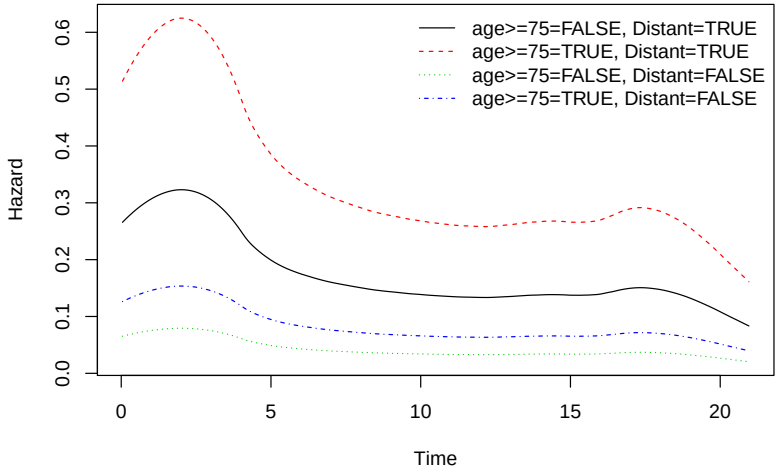
Fitted hazards on the log scale



Cox model adjusted for age at diagnosis

```
> fit <- coxph(Surv(surv_mm,dead) ~ I(age>=75) + I(stage=="Distant"), data=colon2)
> summary(fit)
Call:
coxph(formula = Surv(surv_mm, dead) ~ I(age >= 75) + I(stage == "Distant"), data = colon2)

  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

                          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
```

Cox model adjusted for age at diagnosis (2)



The Cox proportional hazards model (in detail)

- The most commonly applied model in medical time-to-event studies is the Cox proportional hazards model [2].
- 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 survival varies as a function of explanatory variables (the relative rates) rather than the shape of the underlying hazard function (the absolute rate).
- In most statistical models in epidemiology (e.g. linear regression, logistic 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 Cox proportional hazards model (in detail) (2)

- 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.
- We are interested in modelling the hazard function, $\lambda(t; \boldsymbol{X})$, for an individual with covariate vector $\boldsymbol{X}$, where $\boldsymbol{X}$ represents $X_1, \ldots, X_k$.
- The hazard function should be non-negative for all $t > 0$; thus, using

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

may be inappropriate since we cannot guarantee that the linear predictor is always non-negative for all choices of $X_1, \ldots, X_k$ and $\beta_0, \ldots, \beta_k$.

The Cox proportional hazards model (in detail) (3)

- However, $\exp(\beta_0 + \beta_1 X_1 + \cdots + \beta_k X_k)$ is always positive so another option would be

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

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

- In this formulation, both the left and right hand side of the equation can assume any value, positive or negative.
- 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 Cox proportional hazards model (in detail) (4)

- 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 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) (5)

- This regression model for the hazard rate was first introduced by Cox [2], 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.

The Cox proportional hazards model (in detail) (6)

- As with logistic 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.
- Differences between these three test statistics are indicative of possible problems with the fit of the model.

The Cox proportional hazards model (in detail) (7)

- The assumption of proportional hazards is a strong assumption, and should be tested (see separate lecture).
- Because of the inter-relationship between the hazard function, $\lambda(t)$, and the survivor function, $S(t)$, we can show that the PH regression model is equivalent to specifying that

$$S(t|\boldsymbol{X}) = (S_0(t))^{\exp(\beta_1 X_1 + \cdots + \beta_k X_k)} \tag{3}$$

  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.
- Most software packages, will provide estimates of $S(t)$ based on the fitted proportional hazards model for any specified values of explanatory variables.

Interpreting the estimated coefficients

- The estimated coefficients, β, are log rate ratios. To get the rate ratios we need to exponentiate the coeffecients, $\exp(\beta)$.
- The confidence intervals for the β are on the log scale. The CIs are therefore not symmetric around the rate ratios.
- Conceptually identical to logistic regression, where we modelled the log odds as a linear function of covariates and the parameters were interpreted as log odds ratios.
- Here we are modelling the log rate as a linear function of covariates and the parameters are interpreted as log rate ratios.

Interpreting the estimated coefficients

- 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 multiple linear regression, i.e.,

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

- In linear 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 coefficients (2)

- 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,

$$\begin{aligned} \beta X &= \log \lambda(t|X) - \log \lambda_0(t) \\ &= \log \left(\frac{\lambda(t|X)}{\lambda_0(t)} \right) \end{aligned} \tag{4}$$

Interpreting the estimated coefficients (3)

- From the last slide:

$$\beta X = \log\left(\frac{\lambda(t|X)}{\lambda_0(t)}\right) \tag{5}$$

- Since $\lambda_0(t)$ is the hazard function when $X = 0$,

$$\beta = \log\left(\frac{\lambda(t|X=1)}{\lambda(t|X=0)}\right) \tag{6}$$

- 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 hazard ratio and $\exp(\beta)$ is the hazard ratio.

Interpreting the estimated coefficients (4)

- β is also the difference in log hazards, although a difference in log hazards does not have a practical interpretation.

$$\begin{aligned}\beta &= \log\left(\frac{\lambda(t|X=1)}{\lambda(t|X=0)}\right) \\ &= \log(\lambda(t|X=1)) - \log(\lambda(t|X=0))\end{aligned} \tag{7}$$

Interpreting the estimated coefficients (5)

- 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 $X = 0$. That is, the hazard ratio.
- 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.

Example: Localised colon carcinoma 1975–1994

- 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 study cause-specific mortality (`status=="Dead: cancer"`).

Example: Localised colon carcinoma 1975–1994 (2)

```
> fit1 <- coxph(Surv(surv_mm/12, status=="Dead: cancer") ~
               sex + agegrp + year8594,
               subset=(stage=="Localised"), data=colon)
> summary(fit1)

  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

Concordance= 0.609  (se = 0.007 )
Rsquare= 0.031   (max possible= 0.99 )
Likelihood ratio test= 199.1  on 5 df,   p=0
Wald test            = 198.4  on 5 df,   p=0
Score (logrank) test = 204.2  on 5 df,   p=0
```

- 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 ($P < 0.00005$).

Example: Localised colon carcinoma 1975–1994 (3)

```
> fit1 <- coxph(Surv(surv_mm/12, status=="Dead: cancer") ~
               sex + agegrp + year8594,
               subset=(stage=="Localised"), data=colon)
> summary(fit1)

  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

Concordance= 0.609  (se = 0.007 )
Rsquare= 0.031   (max possible= 0.99 )
Likelihood ratio test= 199.1  on 5 df,   p=0
Wald test            = 198.4  on 5 df,   p=0
Score (logrank) test = 204.2  on 5 df,   p=0
```

- The variable sex is coded as 1 for males and 2 for females. Since each parameter represents the effect of a one unit increase in the corresponding variable, the estimated hazard ratio for sex represents the ratio of the hazards for females compared to males.
- That is, the estimated hazard ratio is 0.91 indicating that females have an estimated 9% lower colon cancer mortality than males. There is some evidence that the difference is statistically significant ($P = 0.07$).

Example: Localised colon carcinoma 1975–1994 (4)

```
> fit1 <- coxph(Surv(surv_mm/12, status=="Dead: cancer") ~
               sex + agegrp + year8594,
               subset=(stage=="Localised"), data=colon)
> summary(fit1)

  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

Concordance= 0.609  (se = 0.007 )
Rsquare= 0.031   (max possible= 0.99 )
Likelihood ratio test= 199.1  on 5 df,   p=0
Wald test            = 198.4  on 5 df,   p=0
Score (logrank) test = 204.2  on 5 df,   p=0
```

- The model assumes that the estimated hazard ratio of 0.91 is the same at each and every point during follow-up and for all combinations of the other covariates.
- 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+).

Example: Localised colon carcinoma 1975–1994 (5)

```
> fit1 <- coxph(Surv(surv_mm/12, status=="Dead: cancer") ~
               sex + agegrp + year8594,
               subset=(stage=="Localised"), data=colon)
> summary(fit1)

  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

Concordance= 0.609  (se = 0.007 )
Rsquare= 0.031   (max possible= 0.99 )
Likelihood ratio test= 199.1  on 5 df,   p=0
Wald test            = 198.4  on 5 df,   p=0
Score (logrank) test = 204.2  on 5 df,   p=0
```

- The indicator variable year8594 has the value 1 for patients diagnosed during 1985–1994 and 0 for patients diagnosed during 1975–1984.
- 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 ($P < 0.0005$).

Example: Localised colon carcinoma 1975–1994 (6)

```
> fit1 <- coxph(Surv(surv_mm/12, status=="Dead: cancer") ~
               sex + agegrp + year8594,
               subset=(stage=="Localised"), data=colon)
> summary(fit1)

  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

Concordance= 0.609  (se = 0.007 )
Rsquare= 0.031   (max possible= 0.99 )
Likelihood ratio test= 199.1  on 5 df,   p=0
Wald test            = 198.4  on 5 df,   p=0
Score (logrank) test = 204.2  on 5 df,   p=0
```

- We chose to group age at diagnosis into four categories; 0–44, 45–59, 60–74, and 75+ years.
- It is estimated that individuals aged 75+ at diagnosis experience 2.26 times higher risk of death due to colon carcinoma than individuals aged 0–44 at diagnosis, a difference which is statistically significant.
- 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.

Example: Localised colon carcinoma 1975–1994 (7)

```
> fit1 <- coxph(Surv(surv_mm/12, status=="Dead: cancer") ~
               sex + agegrp + year8594,
               subset=(stage=="Localised"), data=colon)
> summary(fit1)

  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

Concordance= 0.609  (se = 0.007 )
Rsquare= 0.031   (max possible= 0.99 )
Likelihood ratio test= 199.1  on 5 df,   p=0
Wald test            = 198.4  on 5 df,   p=0
Score (logrank) test = 204.2  on 5 df,   p=0
```

- 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.

Example: Localised colon carcinoma 1975–1994 (8)

```
> library(car)
> linearHypothesis(fit1, 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
```

- 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.

Example: Localised colon carcinoma 1975–1994 (9)

```
> fit2 <- coxph(Surv(surv_mm/12, status=="Dead: cancer")
                ~ sex + year8594,
                subset=(stage=="Localised"), data=colon)
> anova(fit1,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 Wald test is an approximation to the likelihood ratio test, which compares the likelihood between models.
- To perform a likelihood ratio test we fit the reduced model (the model without age) and see that the log likelihood is -14430.

Example: Localised colon carcinoma 1975–1994 (10)

```
> fit2 <- coxph(Surv(surv_mm/12, status=="Dead: cancer")
                ~ sex + year8594,
                subset=(stage=="Localised"), data=colon)
> anova(fit1,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

```
> fit3 <- coxph(Surv(surv_mm/12, status=="Dead: cancer") ~ sex + age + year8594,
               subset=(stage=="Localised"), data=colon)
> summary(fit3)

  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 ***
```

- 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 $1.034^{10} = 1.40$.

References

[1] Danehy TJ, Lock RH. CHODR – Using statistics to predict college hockey. *Stats The Magazine for Students of Statistics* 1995;**13**:10–14.

[2] Cox DR. Regression models and life-tables (with discussion). *JRSSB* 1972;**34**:187–220.