Applied Statistical Modeling 2

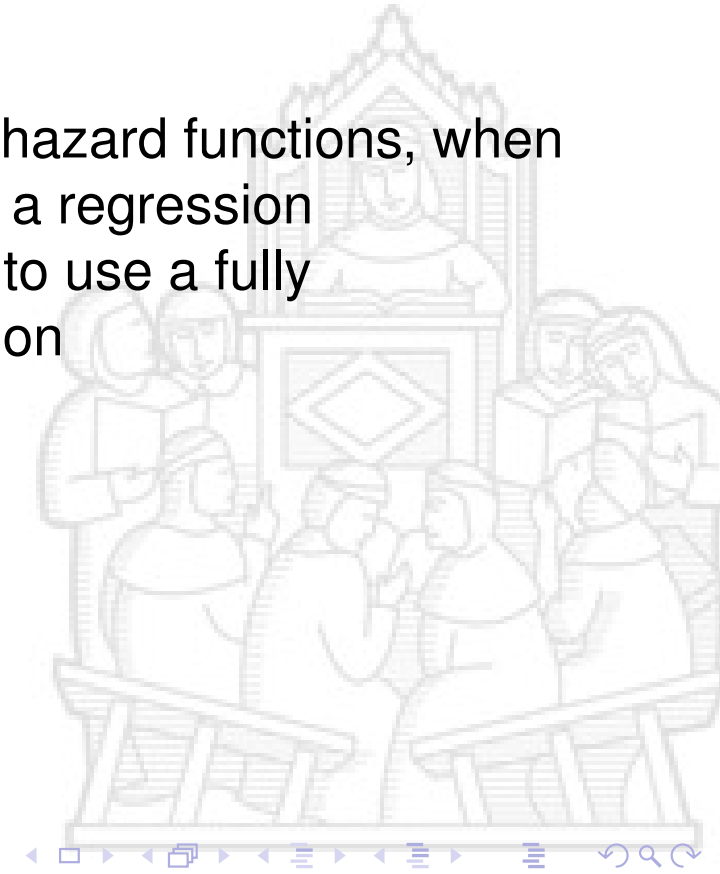
Lorenzoni Valentina
valentina.lorenzoni@santannapisa.it

Institute of Management - Scuola Superiore Sant' Anna, Pisa

21st May 2025

Regression models for survival data

Rather than simply describing the surviva and hazard functions, when dealing with survival data the interest is to use a regression model-like structure in those functions, that is, to use a fully parametrization of the survival or hazard function

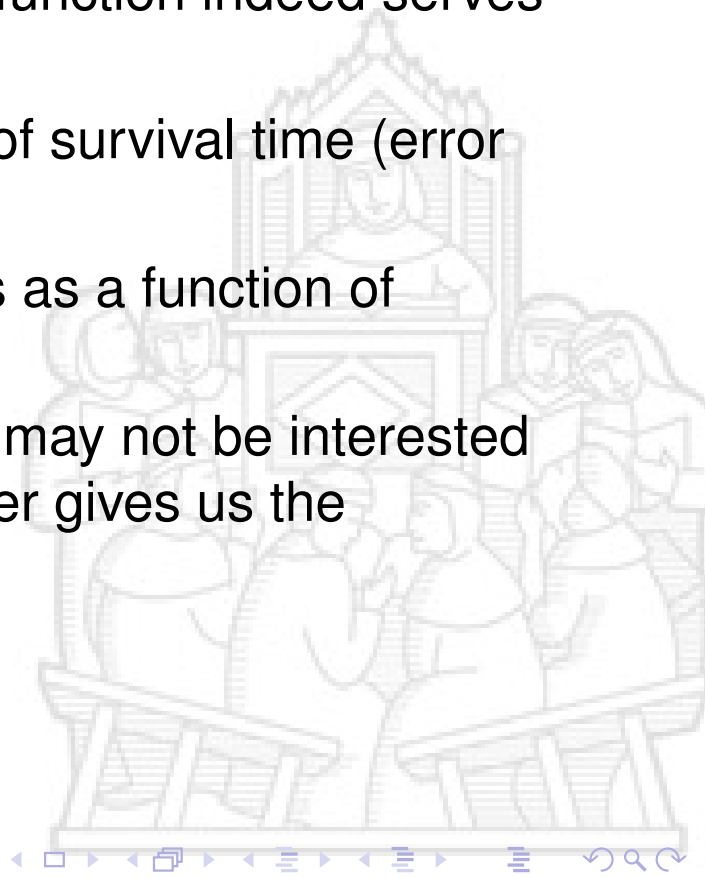


Regression models for survival data

A full parametrization of the survival or hazard function indeed serves to:

- Describe the basic underlying distribution of survival time (error component), and
- Characterize how that distribution changes as a function of covariates (systematic component)

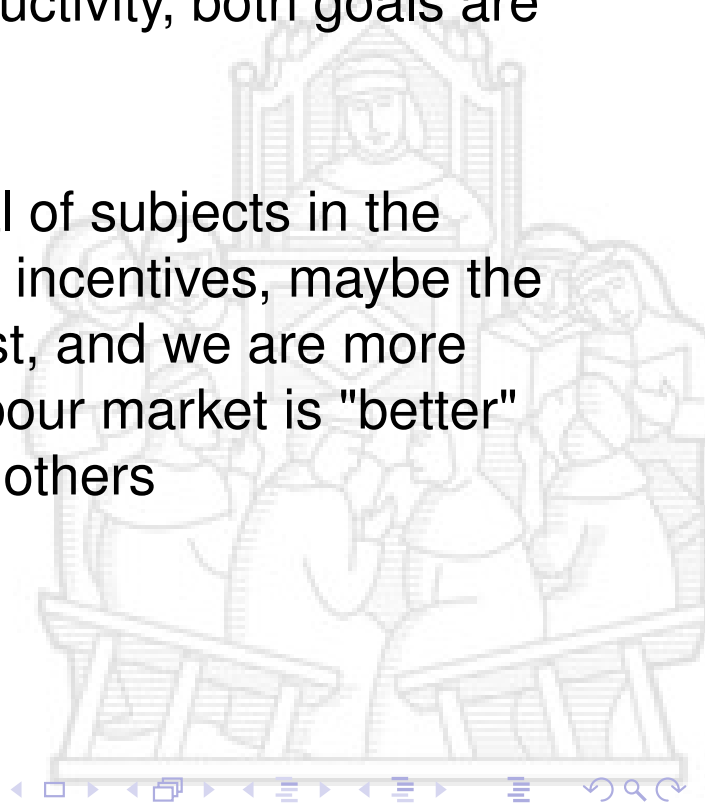
both goals are relevant, but in some cases, we may not be interested in both, and maybe, in some situations, the latter gives us the necessary information we are interested in



Regression models for survival data?

If we are interested in understanding the survival time of firms in a particular market as a function of age and productivity, both goals are essential

While, if we are interested in evaluating survival of subjects in the labour market according to two different sets of incentives, maybe the description of the survival time is less of interest, and we are more interested in understanding if survival in the labour market is "better" with a certain set of incentives as compared to others

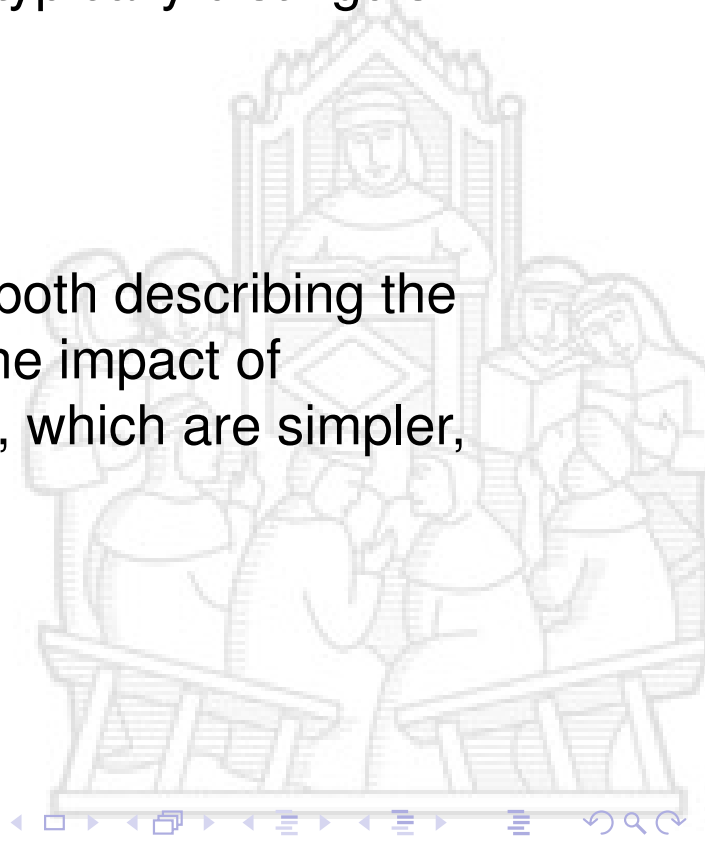


Regression models for survival data?

Accordingly, and depending on their goals, we typically distinguish among:

- fully **parametric regression models**, and
- **semi-parametric regression models**

while fully parametric regression models allow both describing the basic underlying survival time and estimating the impact of covariates, semi-parametric regression models, which are simpler, enable just the latter



Semi-parametric regression models for survival data

Let's think about

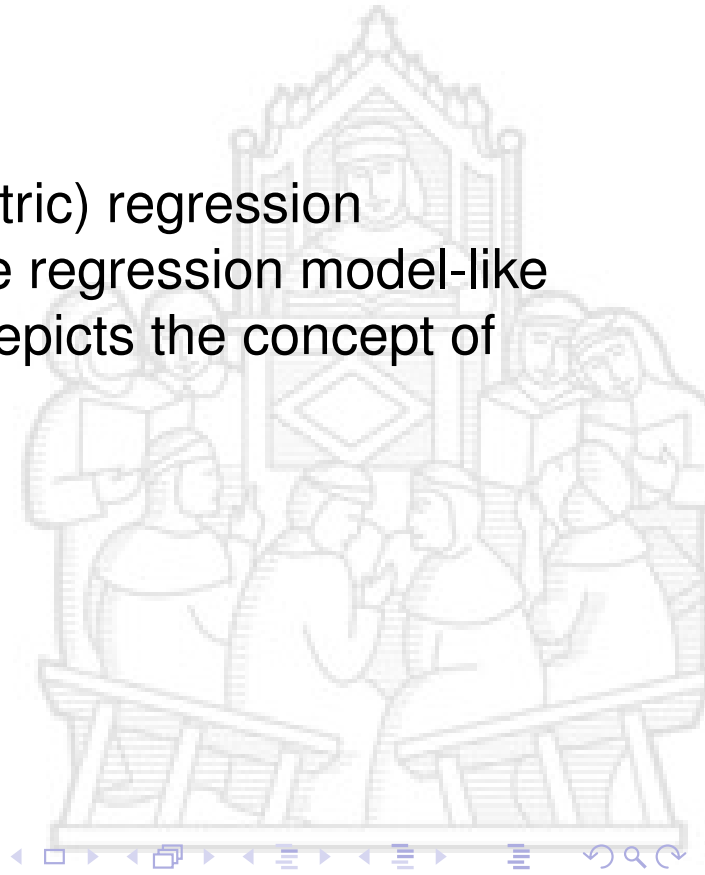
- being interested in comparing the survival experience in the labour market of two groups of subjects, each targeted with different sets of incentives, or even
- wishing to compare the survival experience of drug addicted individuals treated in two different clinics (as done descriptively for the *addicts* dataset)

Semi-parametric regression models for survival data

In those circumstances we could use semi-parametric regression models to characterize the distribution of the survival function according to the covariate of interest, without "putting effort" on describing the basic underlying distribution of survival time

Semi-parametric regression models for survival data

When fitting (fully parametric and semi-parametric) regression models to survival data, we generally use some regression model-like structure for the hazard function as it "better" depicts the concept of "risk" over time

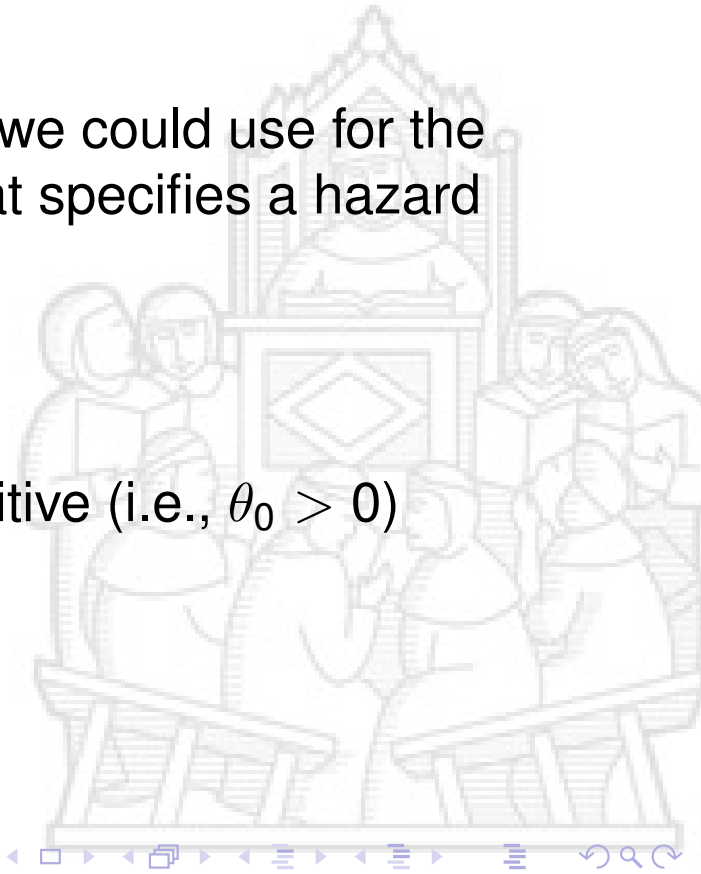


Semi-parametric regression models for survival data

Generally speaking, there are different models we could use for the hazard function, the simplest one is the one that specifies a hazard function that is constant over time

$$h(t) = \theta_0$$

as the hazard function is a rate, it is strictly positive (i.e., $\theta_0 > 0$)



Semi-parametric regression models for survival data

To allow the possibility that the range of parameter values is infinite (a desirable property that a statistical model should have), the hazard function could be better parametrized using the following

$$h(t) = e^{\beta_0}$$

were $\beta_0 = ln(\theta_0)$

Semi-parametric regression models for survival data

Using that parametrization, we could add covariates using an additive form in the log scale

$$h(t, \beta) = e^{\beta_0 + \beta_1 x}$$

so that

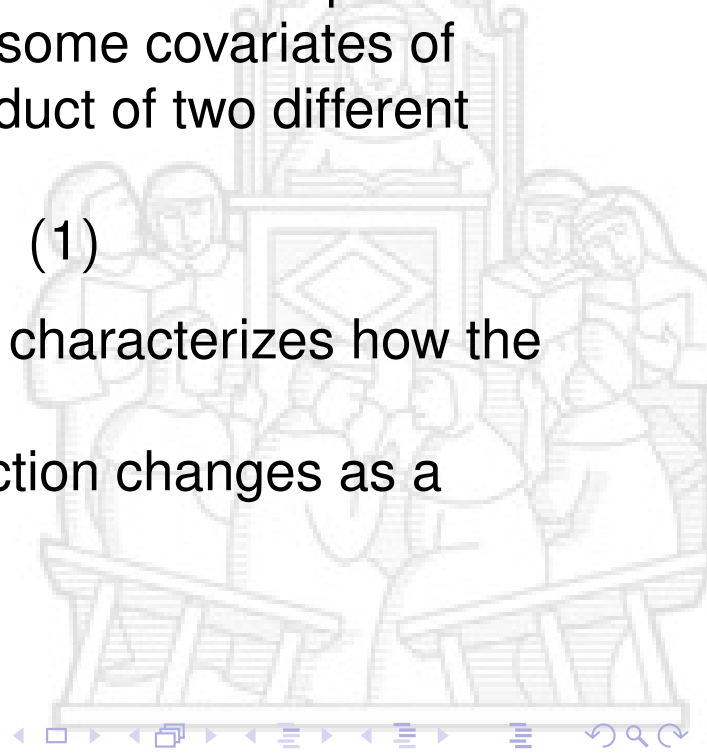
$$ln[h(t, \beta)] = \beta_0 + \beta_1 x$$

Semi-parametric regression models for survival data

Considering the parametrization introduced before, a semi-parametric regression model for the hazard function that addresses the problem of comparing survival experience according to some covariates of interest could be expressed in terms of the product of two different functions

$$h(t, x, \beta) = h_0(t) r(x, \beta) \qquad (1)$$

where h_0 is the **baseline hazard function** that characterizes how the hazard function changes as a function of time,
while $r(x, \beta)$ characterizes how the hazard function changes as a function of individual covariates



Semi-parametric regression models for survival data

Remembering that the hazard function is a rate, so strictly positive, the two functions need to be chosen so that

$$h(t, x, \beta) > 0$$

and when the function is parametrized such that

$$r(x = 0, \beta) = 1$$

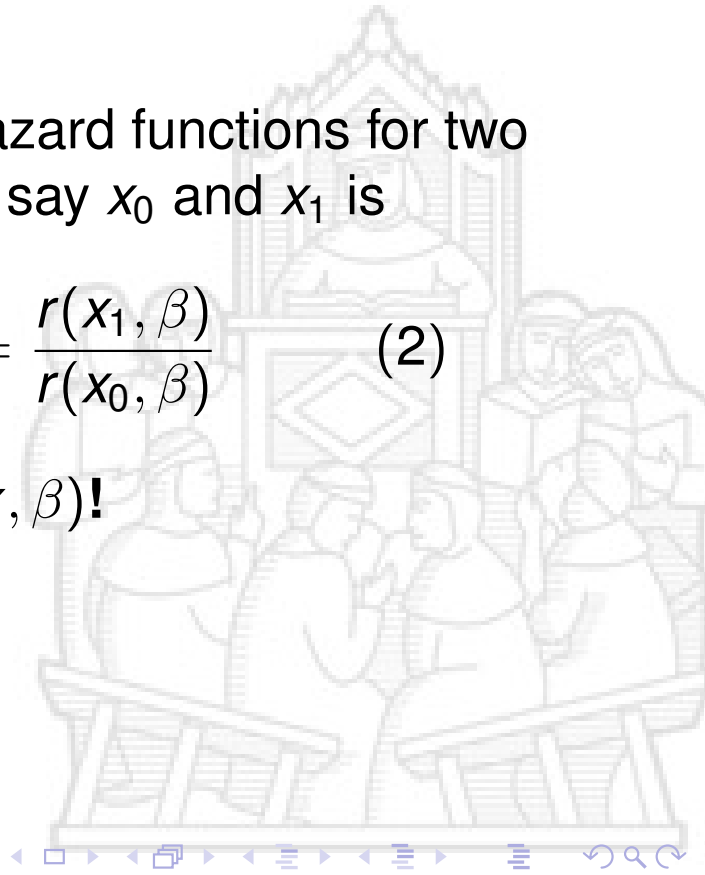
$h_0(t)$ could be thought as a generalization of the intercept in a parametric regression model

Semi-parametric regression models for survival data

Considering the model in (1), the ratio of the hazard functions for two subjects with different values of the covariates, say x_0 and x_1 is

$$HR(t, x_1, x_0) = \frac{h(t, x_1, \beta)}{h(t, x_0, \beta)} = \frac{h_0(t) r(x_1, \beta)}{h_0(t) r(x_0, \beta)} = \frac{r(x_1, \beta)}{r(x_0, \beta)} \qquad (2)$$

and **does not depend on time but just on** $r(x, \beta)$**!**



Cox proportional hazard model

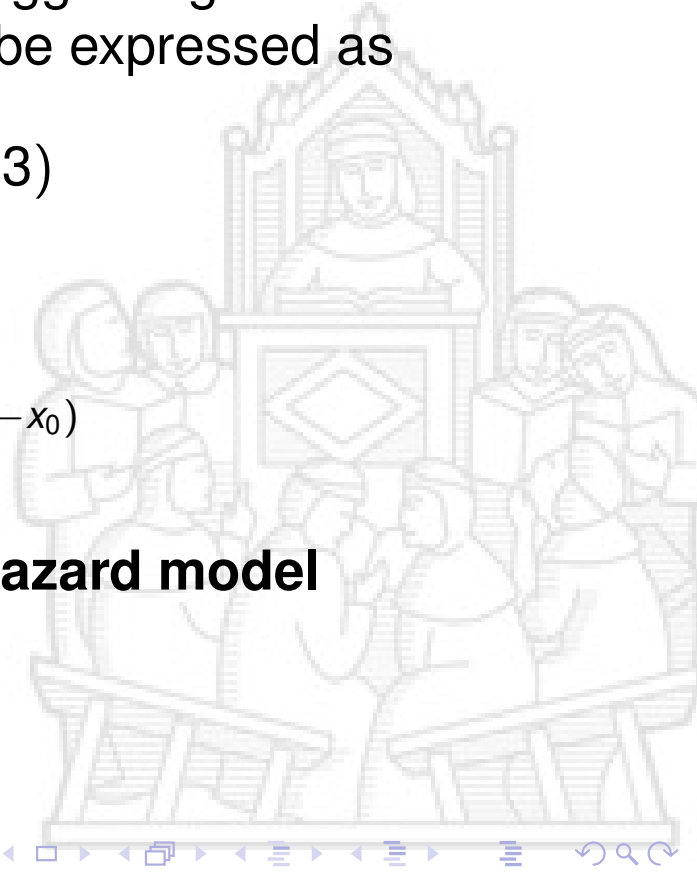
Cox D (1972) first proposed the model in (2), suggesting the use of $r(x, \beta) = e^{x\beta}$ so that the hazard function could be expressed as

$$h(t, x, \beta) = h_0(t)e^{x\beta} \qquad (3)$$

and the hazard ratio becomes

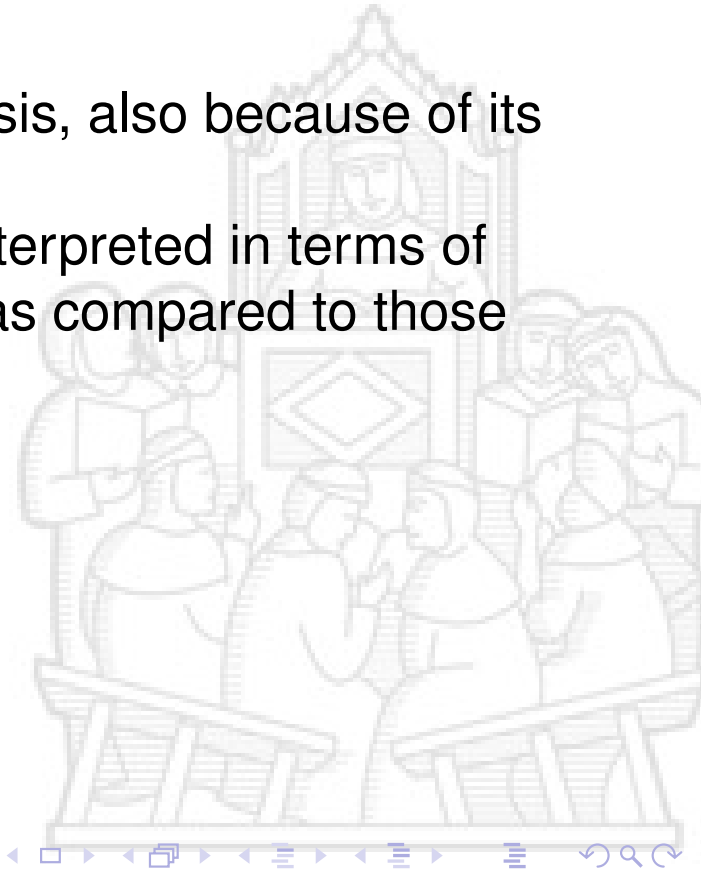
$$HR(t, x_1, x_0) = \frac{e^{x_1,\beta}}{e^{x_0,\beta}} = e^{\beta(x_1 - x_0)}$$

which is well known as the **Cox proportional hazard model**



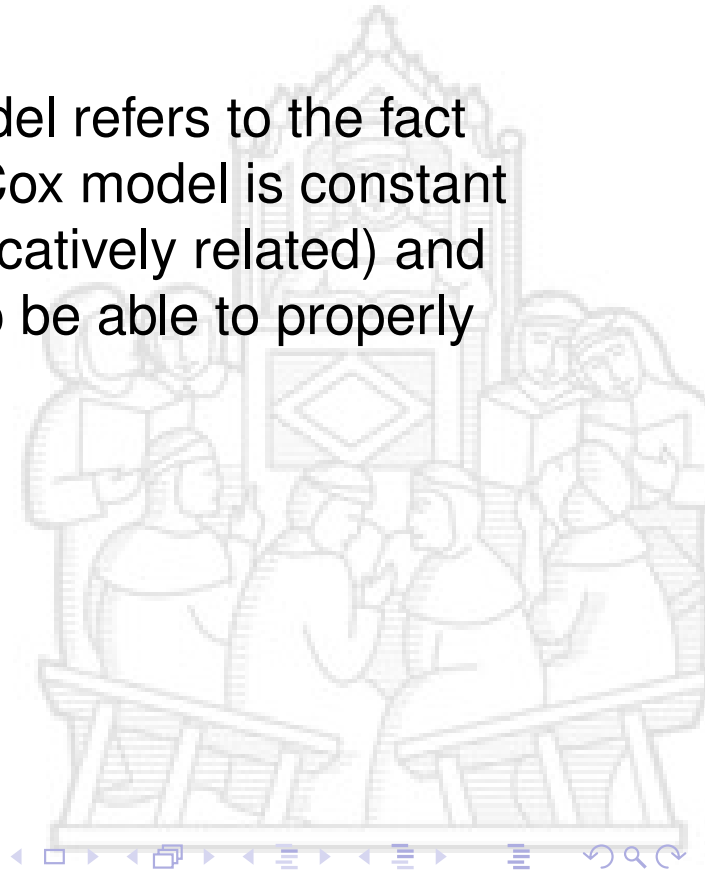
Cox proportional hazard model

The Cox model is widely used in survival analysis, also because of its ease of interpretation in terms of "relative risk"
Indeed coefficients from that model could be interpreted in terms of the relative risk for subjects with covariates x_1 as compared to those with covariate values x_0



Cox proportional hazard model

The term **proportional hazard** for the Cox model refers to the fact that the hazard ratio that is obtained from the Cox model is constant over time (i.e., the hazard functions are multiplicatively related) and that assumption needs to be verified in order to be able to properly use the Cox model



Cox proportional hazard model

As the survival function could be expressed using the cumulative hazard function H(t) as

$$S(t) = e^{-H(t)}$$

and adding covariates

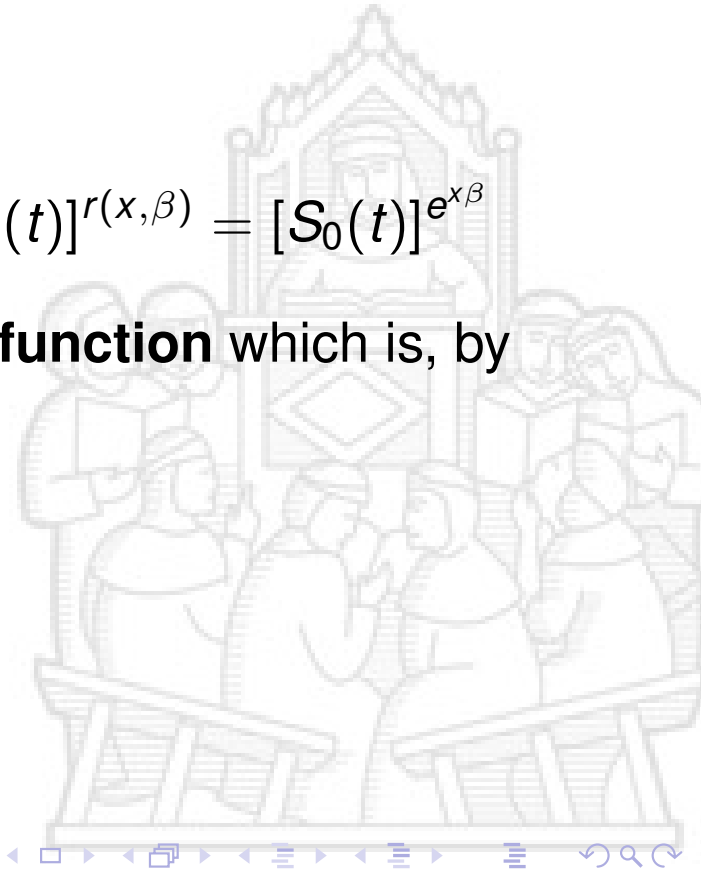
$$S(t, x, \beta) = e^{-H(t,x,\beta)}$$

Cox proportional hazard model

Using the parametrization proposed by Cox

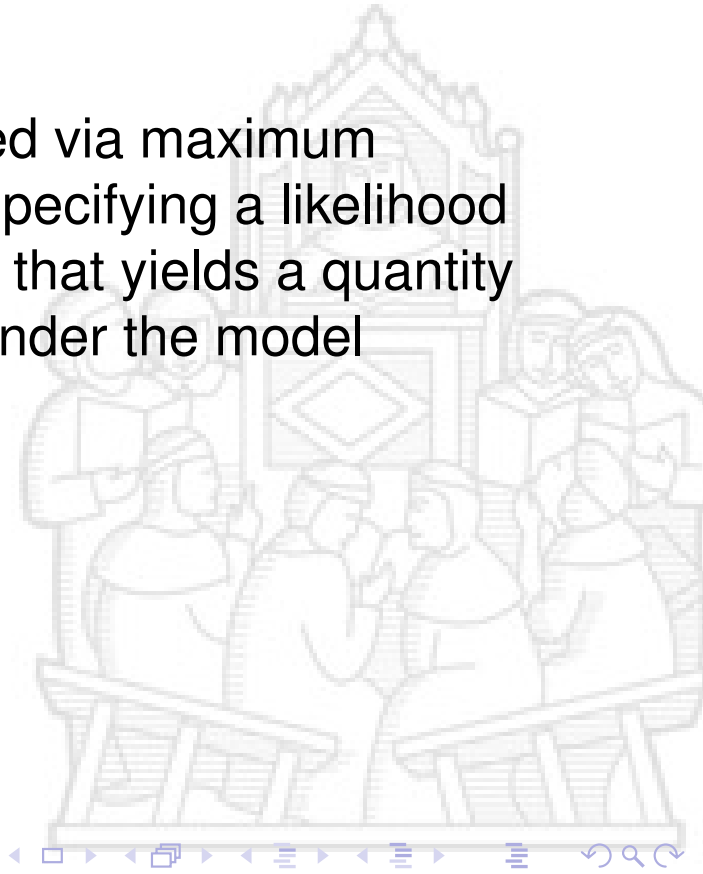
$$S(t, x, \beta) = e^{-r(x,\beta)H_0(t)} = [e^{-H_0(t)}]^{r(x,\beta)} = [S_0(t)]^{r(x,\beta)} = [S_0(t)]^{e^{x\beta}}$$

where $S_0(t) = e_{-H_0(t)}$ is the **baseline survival function** which is, by definition, constrained in the interval [0,1]



Cox proportional hazard model

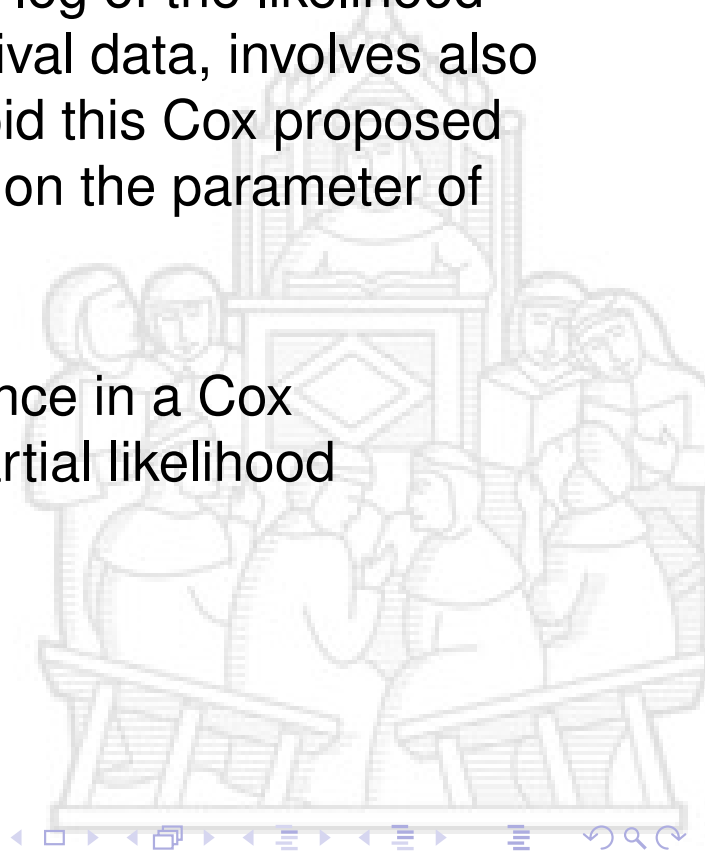
Parameter estimation in a Cox model is obtained via maximum likelihood estimation (MLE), which consists in specifying a likelihood function to be maximized, that is an expression that yields a quantity similar to the probability of the observed data under the model



Cox proportional hazard model

Solution of the MLE is found by maximizing the log of the likelihood function, $L(\beta)$, with respect to β which, for survival data, involves also the specification of the error component, to avoid this Cox proposed the partial likelihood function that just depends on the parameter of interest, β

Estimates of model coefficients and of its variance in a Cox regression models are thus obtained via the partial likelihood estimator



Cox proportional hazard model

Overall significance of the model coefficient could be done using:

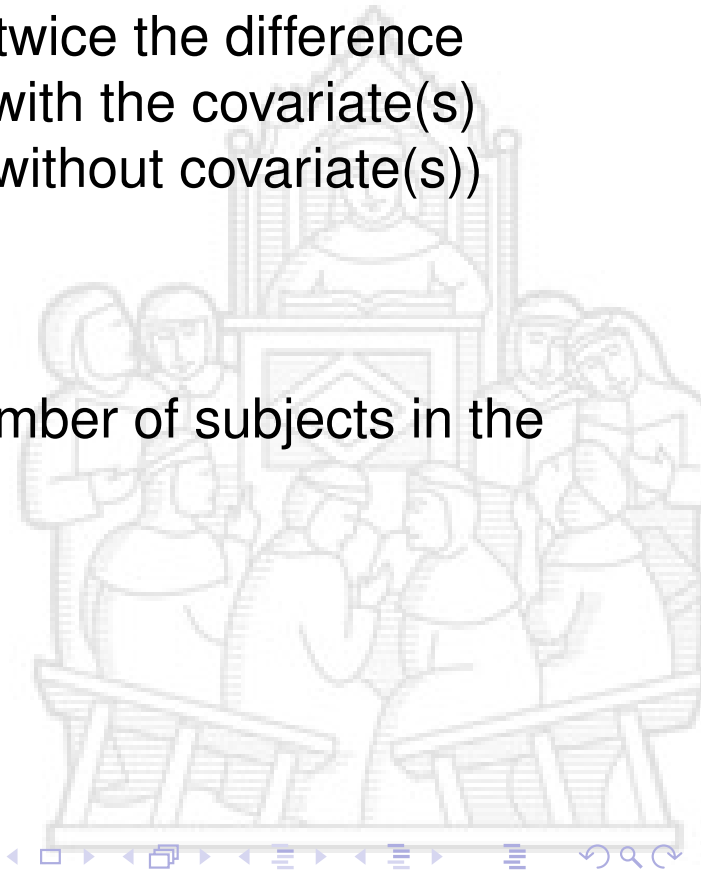
- the partial likelihood ratio test
- the Wald test

Cox proportional hazard model

The partial likelihood ratio test is calculated as twice the difference between the log partial likelihood of the model with the covariate(s) and the log partial likelihood of the null model (without covariate(s))

$$G = 2[L_p(\hat{\beta}) - L_p(\hat{0})]$$

where $L_p(\hat{0}) = \sum ln(n_i)$ with n_i indicates the number of subjects in the risk set at observed survival time $t_{(i)}$



Cox proportional hazard model

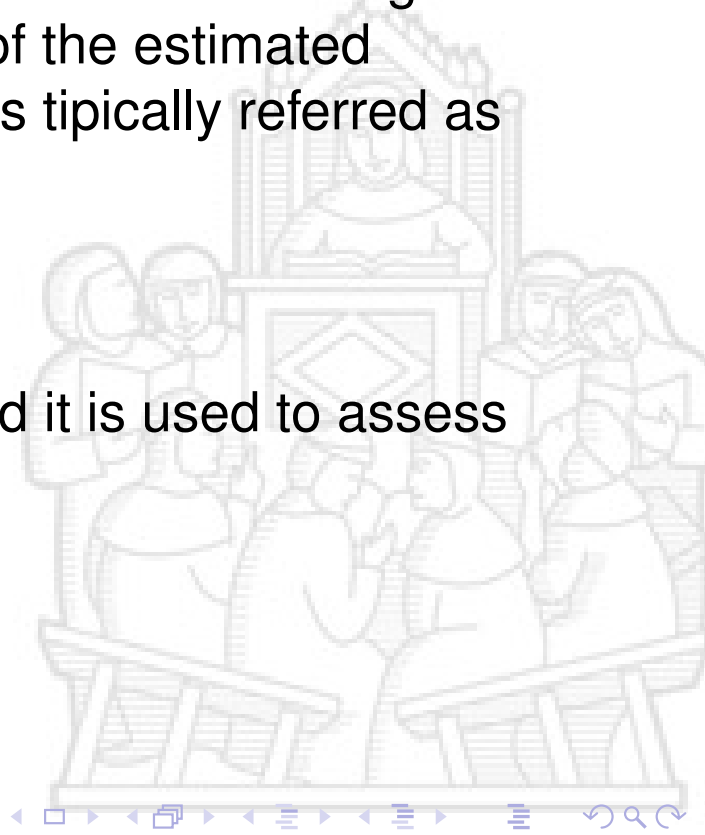
The G statistic follows a chi-square distribution with 1 degree of freedom and could be used to assess overal significance of coefficients

Cox proportional hazard model

The statistic that is generally offered from statistical software together with estimates of model coefficient is the ratio of the estimated coefficient to its estimated standard error, that is tipically referred as the Wald statistic

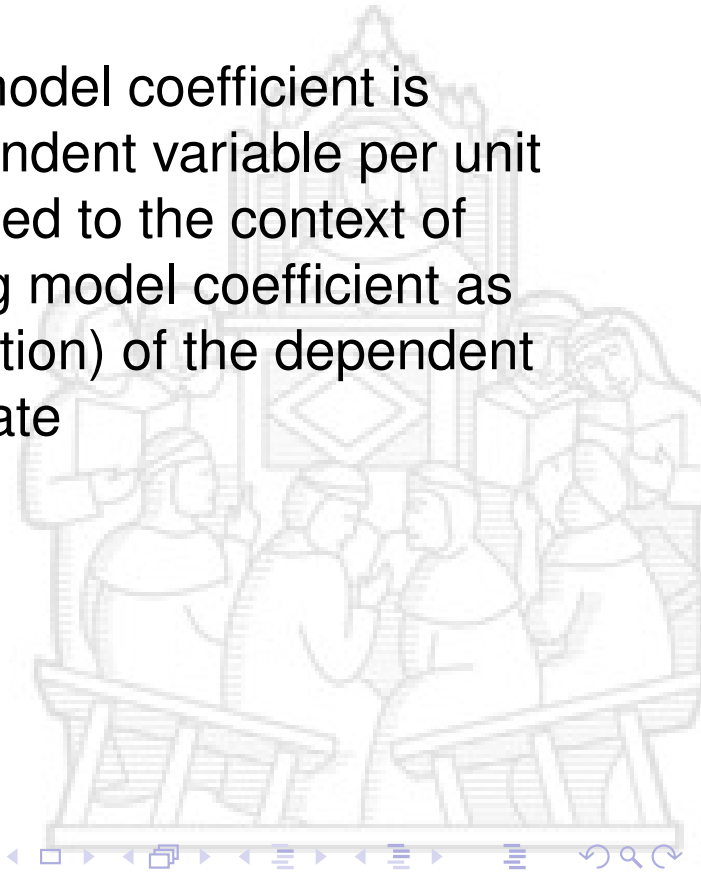
$$z = \frac{\hat{\beta}}{\hat{SE}(\hat{\beta})}$$

which follows a standard normal distribution and it is used to assess the significance of the single coefficient



Cox proportional hazard model

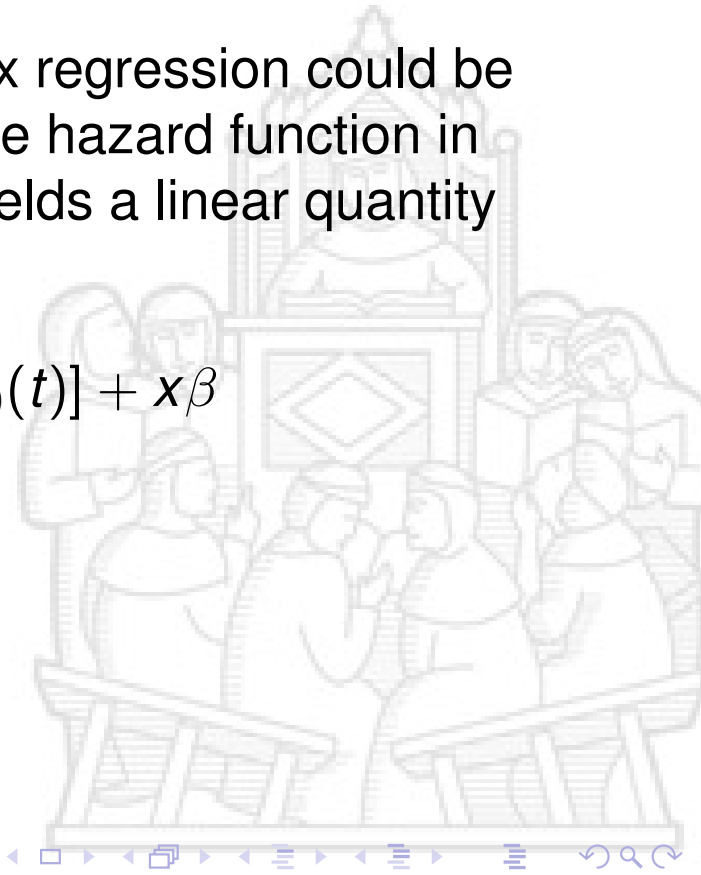
In a linear regression model, interpretation of model coefficient is done in terms of the rate of change of the dependent variable per unit change of the covariate, this could be generalized to the context of the *generalized linear model* (GLM) interpreting model coefficient as the change in a *function* (the so colled link function) of the dependent variable induced by a unit change of the covariate



Cox proportional hazard model

The interpretation of model coefficients in a Cox regression could be done considering that taking the logarithm of the hazard function in the case of a Cox proportional hazard model yields a linear quantity in the parameters

$$ln[h(t, x, \beta)] = ln[h_0(t)e^{x\beta}] = ln[h_0(t)] + x\beta$$



Cox proportional hazard model

Thus the difference in the log hazard for a change from $x = a$ and $x = b$ is

$$ln[h(t, x = a, \beta)] - ln[h(t, x = b, \beta)] = [ln[h_0(t)] + a\beta] - [ln[h_0(t)] + b\beta]$$

$$= a\beta - b\beta = (a - b)\beta$$

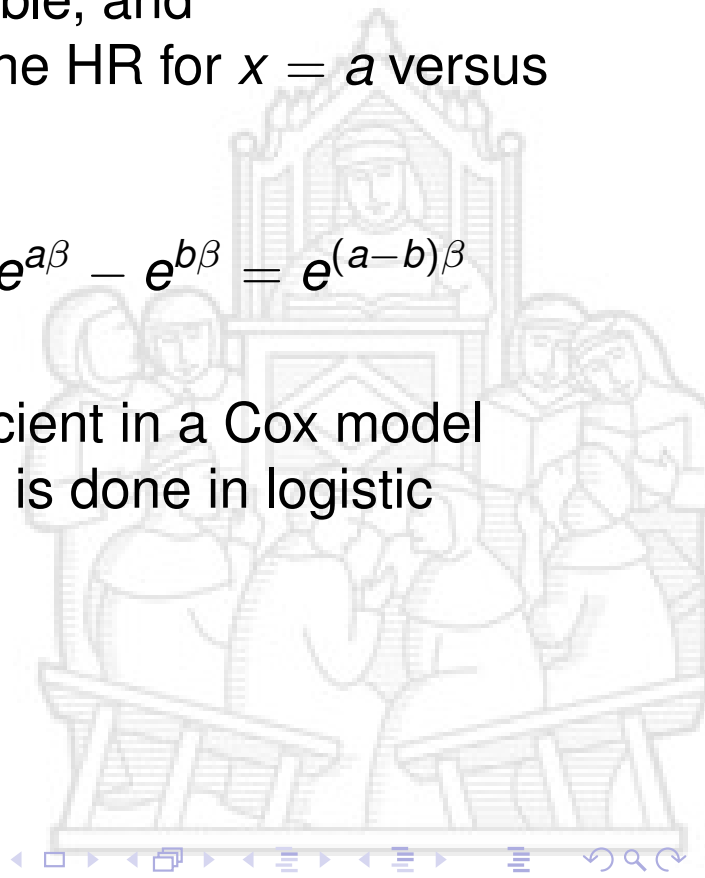
which show that the log hazard is the proper function to assess the effect of change in a covariate, moreover the equation above shows that in the case of a Cox proportional hazard model the difference in the log hazard does not depend on time (the proportional hazard assumption!)

Cox proportional hazard model

However, the log hazard is not easily interpretable, and exponentiating the equation above, we obtain the HR for $x = a$ versus $x = b$

$$HR(t, x = a\, vs\, x = b, \beta) = \frac{h(t, x = b, \beta)}{h(t, x = a, \beta)} = e^{a\beta} - e^{b\beta} = e^{(a-b)\beta}$$

which shows that interpretation of model coefficient in a Cox model could be done using the HR, "similarly" to what is done in logistic regression with odds ratio



Cox proportional hazard model

Assume to use the Cox model to assess the effect of a single dichotomous covariate (i.e., taking just values equal to 0 or 1), x, on the survival experience of a population of interest

In this setting, the difference in the log hazard corresponding to a unit change in the covariate is

$$ln[h(t,1,\beta)] - ln[h(t,0,\beta)] = [ln[h_0(t)] + 1\beta] - [ln[h_0(t)] + 0\beta] = \beta - 0 = \beta$$

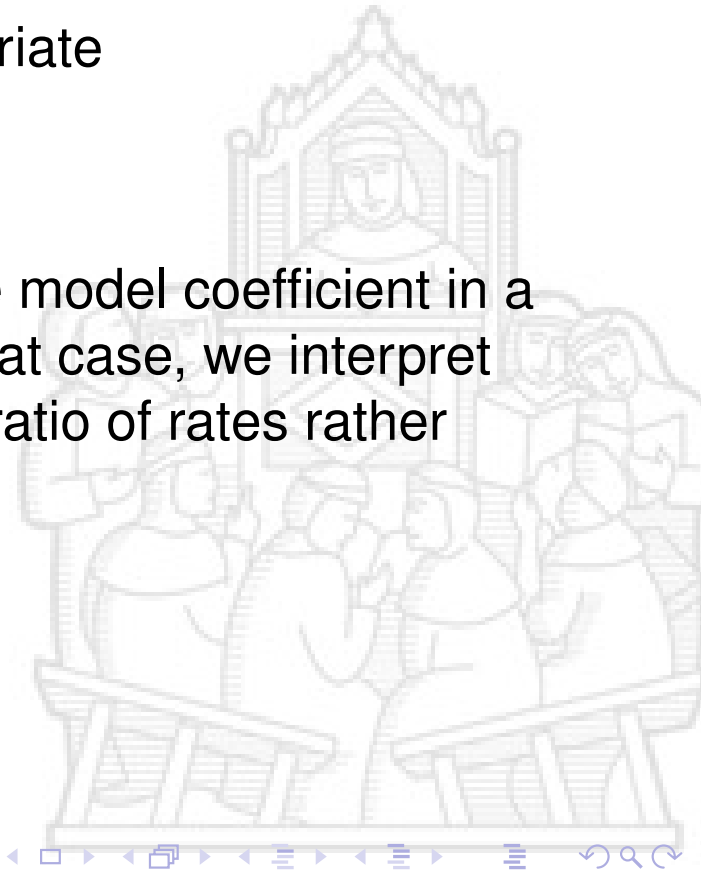
which means that when using a single dichotomous covariate in a Cox model, the coefficient corresponds to the change in the log hazard for x changing from 0 to 1

Cox proportional hazard model

Thus, in the case of a single dichotomous covariate

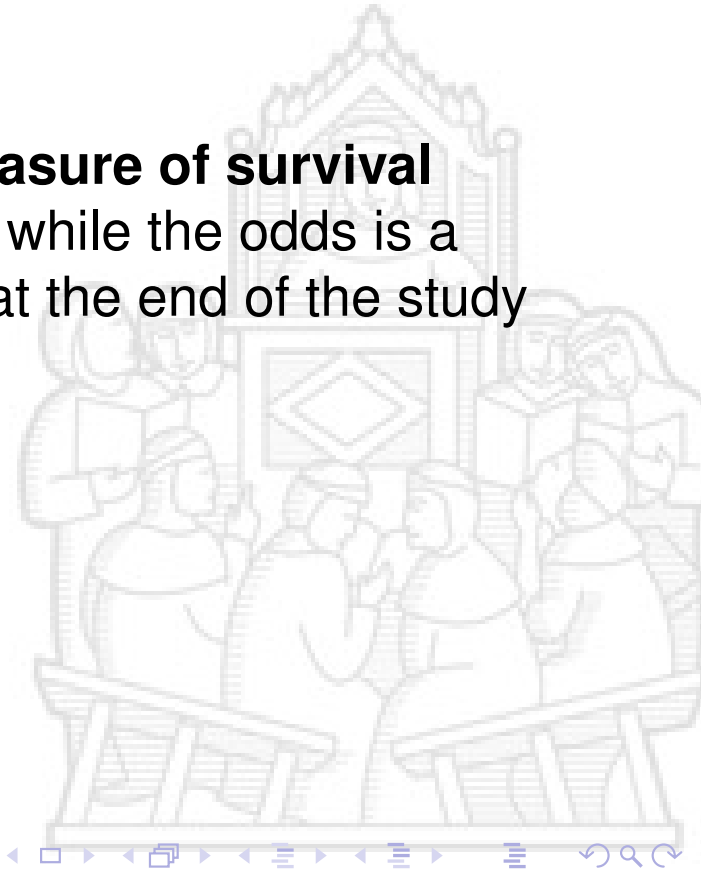
$$HR(t, 1, 0, \beta) = e^{\beta}$$

(again) this remembers the interpretation of the model coefficient in a logistic regression, with the difference that in that case, we interpret the exponentiated coefficient with respect to a ratio of rates rather than a ratio of odds



Cox proportional hazard model

In brief, the **hazard ratio is a comparative measure of survival experience over the entire follow-up period**, while the odds is a comparative measure of the odds of the event at the end of the study

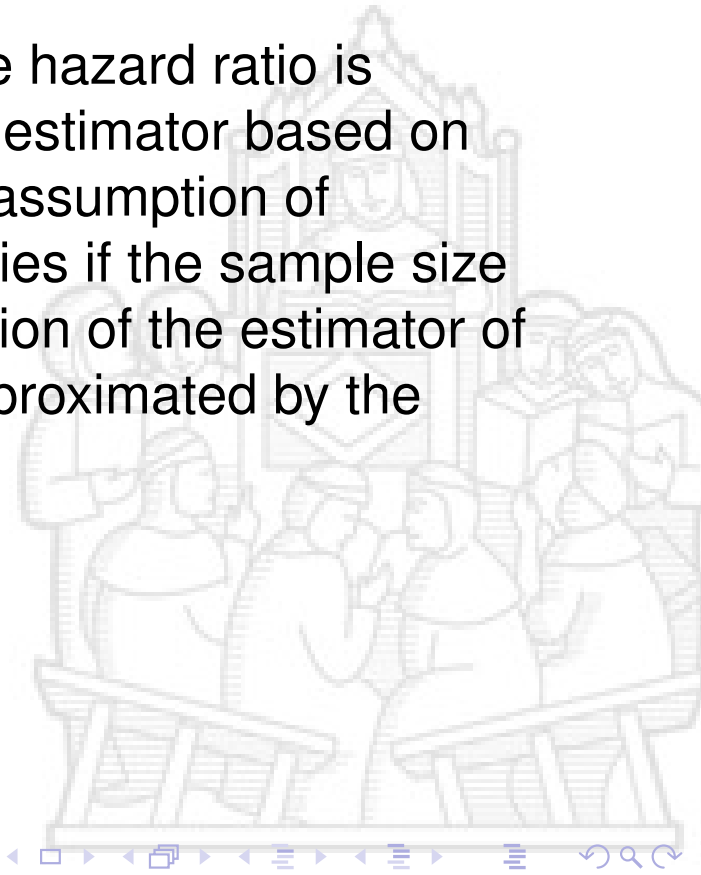


Cox proportional hazard model

As already mentioned partial MLE approach allows obtaining estimates of the model coefficients $\hat{\beta}$ (and their standard error), from which we can easily derive an estimate for the hazard ratio, $\hat{HR}$

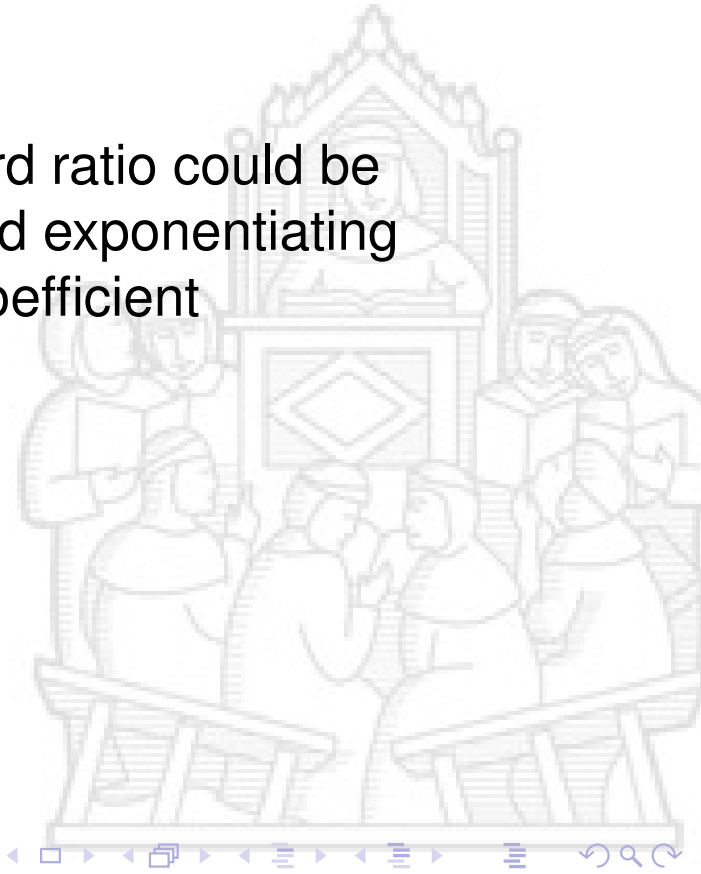
Cox proportional hazard model

The sampling distribution of the estimator of the hazard ratio is skewed to the right and the confidence interval estimator based on the Wald statistic (for the hazard ratio), and its assumption of normality, may not have good coverage properties if the sample size is not large enough while the sampling distribution of the estimator of the (non-exponentiated) coefficient is better approximated by the normal distribution



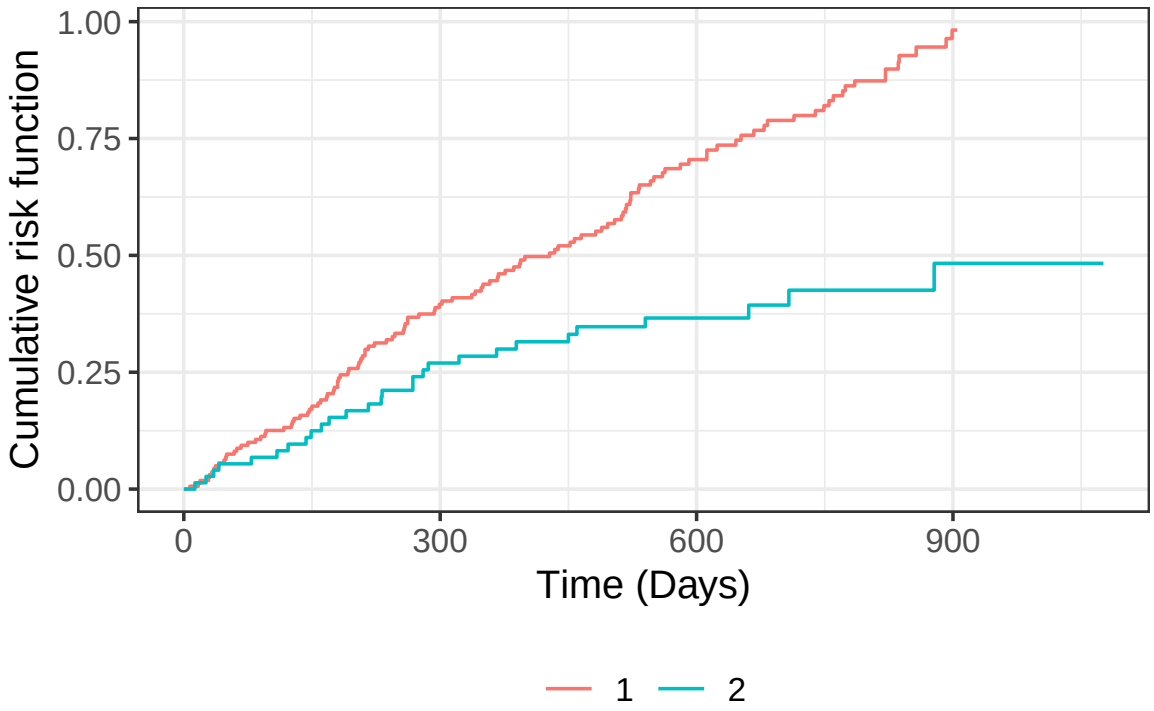
Cox proportional hazard model

Thus, the 95% confidence interval for the hazard ratio could be obtained by using the normal approximation and exponentiating confidence limits for the (non-exponentiated) coefficient



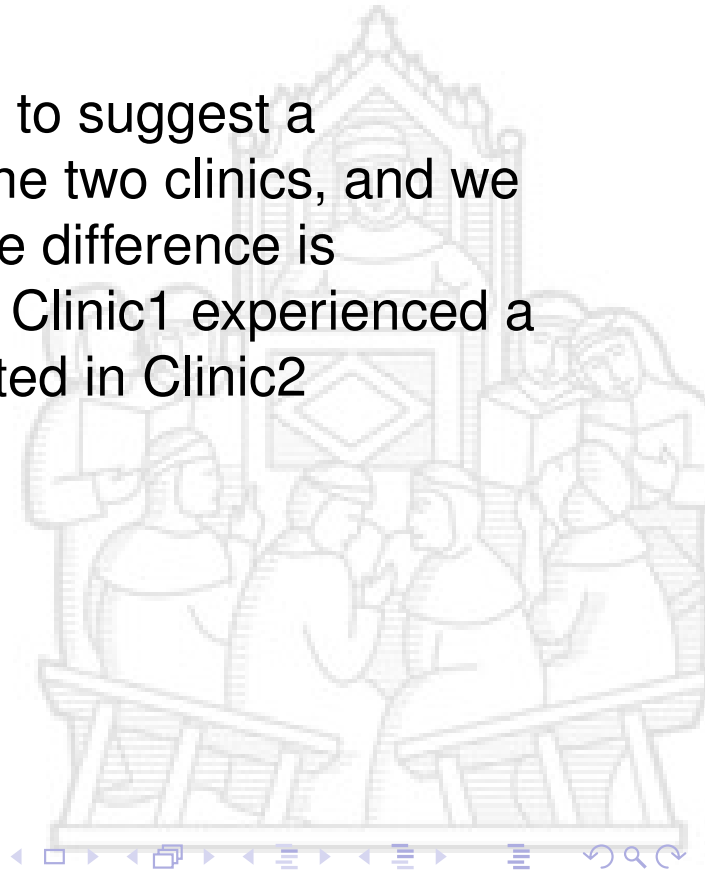
Cox proportional hazard model

Consider the example of the *addicts* dataset and go on with the comparison of survival experience in the two clinics



Cox proportional hazard model

The plot of the cumulative risk functions seems to suggest a difference in the survival experience between the two clinics, and we already verified using the Log-Rank test that the difference is statistically significant, thus, subjects treated in Clinic1 experienced a higher risk of failure as compared to those treated in Clinic2

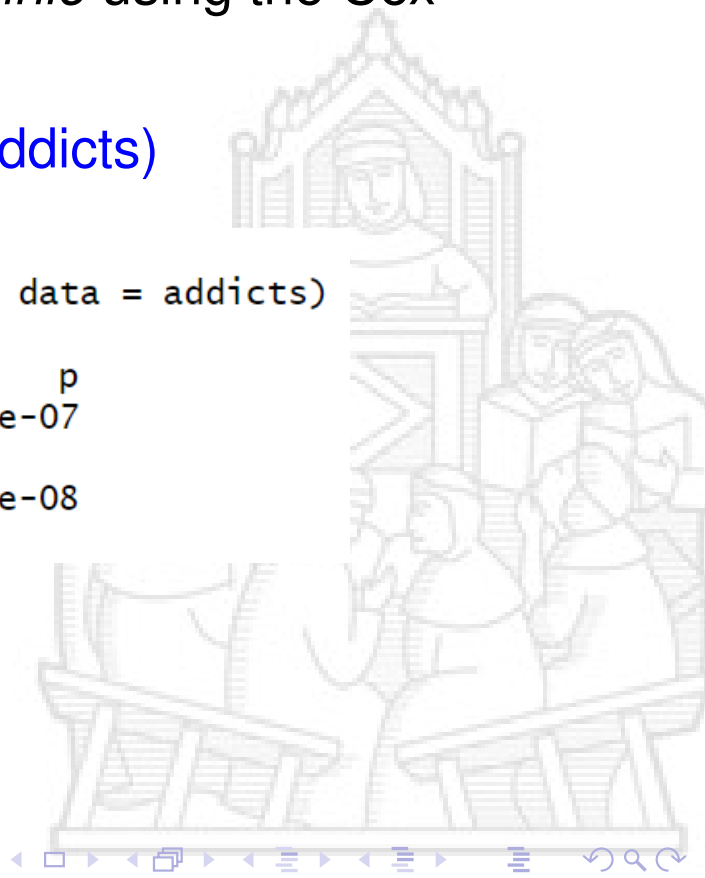


Cox proportional hazard model

Let's try to estimate the effect of the variable *Clinic* using the Cox model, by typing in R

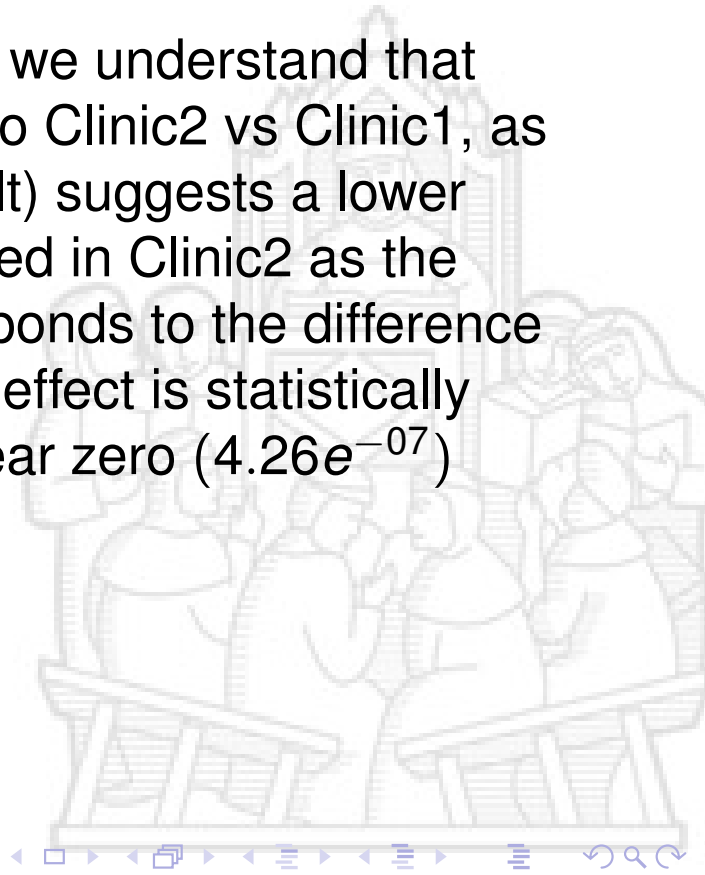
m1<-coxph(Surv(survt, status) clinic, data = addicts)

```
Call:
coxph(formula = Surv(survt, status) ~ clinic, data = addicts)

          coef exp(coef) se(coef)      z        p
clinic -1.0754    0.3412   0.2127 -5.057 4.26e-07

Likelihood ratio test=30.99  on 1 df, p=2.594e-08
n= 238, number of events= 150
```



Cox proportional hazard model

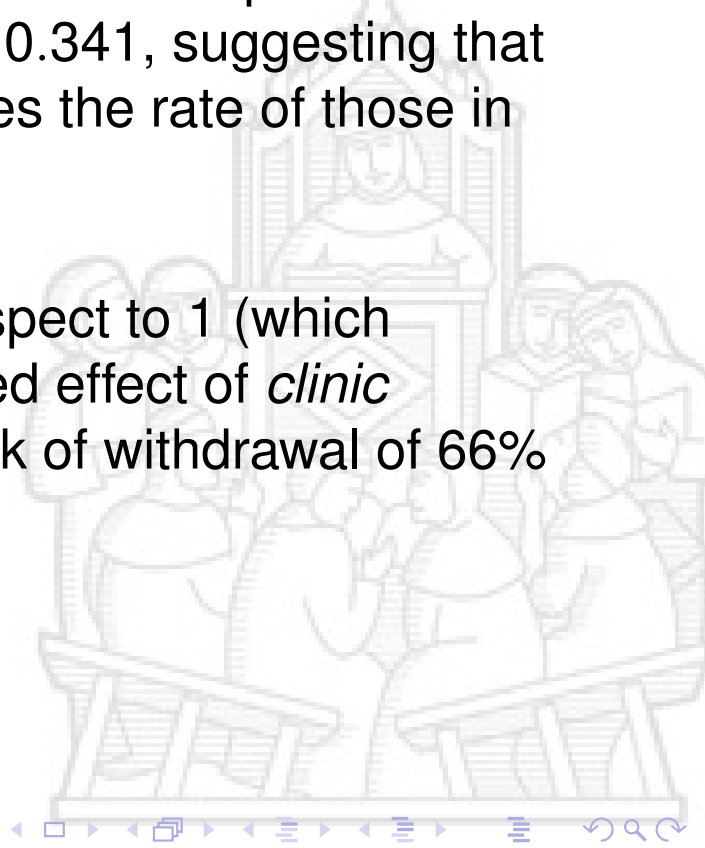
Just focusing for the moment on the estimates, we understand that the estimated coefficient, $\hat{\beta}$, (which is referred to Clinic2 vs Clinic1, as R takes the lowest value as reference by default) suggests a lower risk of withdrawn from treatmnet for those treated in Clinic2 as the estimate is negative, $\hat{\beta} = -1.075$ which corresponds to the difference in the log hazard for Clinic2 versus Clinic1, the effect is statistically significant as the p-value for the Wald test is near zero ($4.26e^{-07}$)



Cox proportional hazard model

Using the exponentiated coefficients which could be interpreted in terms of hazard ratio, we have $exp(\hat{\beta}) = \hat{HR} = 0.341$, suggesting that subjects in Clinic2 withdraw at about 0.341 times the rate of those in Clinic1

Being a ratio, we can also interpret $\hat{HR}$ with respect to 1 (which indicates equal rates) and say that the estimated effect of *clinic* translates into an estimated reduction of the risk of withdrawal of 66% in Clinic2 vs Clinic1 over the study period



Cox proportional hazard model

The 95% confidence interval for $\hat{HR}$ could be obtained typing

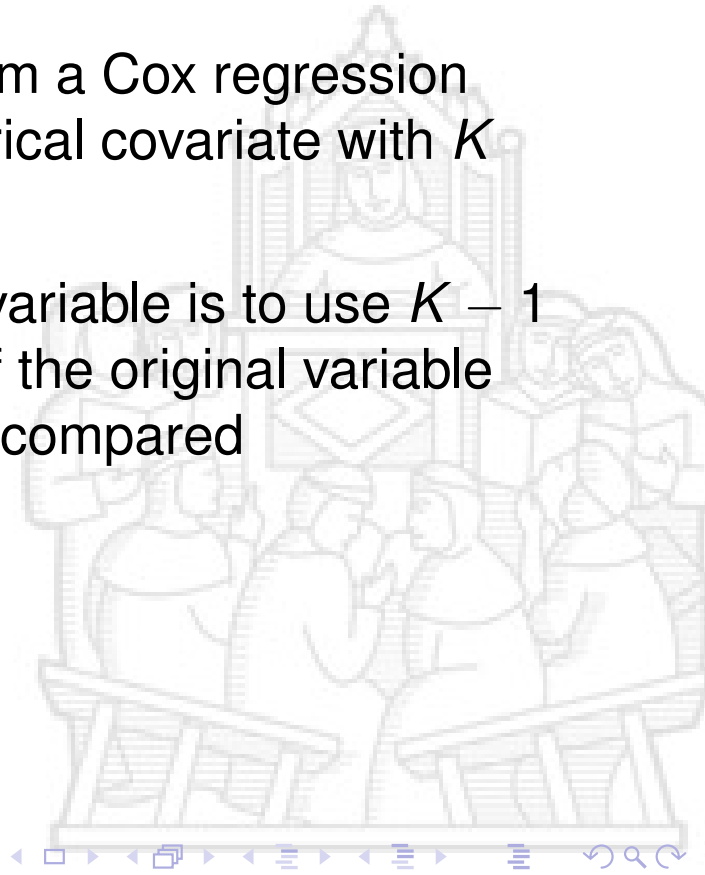
```
exp(confint(m1))
```

which provides a 95% with limits going from 0.225 to .518

Cox proportional hazard model

To broaden the interpretation of coefficients from a Cox regression model, let's consider now the case of a categorical covariate with K distinct values

A commonly used approach to deal with such variable is to use $K - 1$ dummy variables choosing one of the k level of the original variable as reference, against which all other levels are compared



Cox proportional hazard model

Using this approach we can estimate $k - 1$

$$\hat{HR}_k = e^{\hat{\beta}_k}$$

each estimating the hazard ratio for the k^{th} group to that of the reference group

Cox proportional hazard model

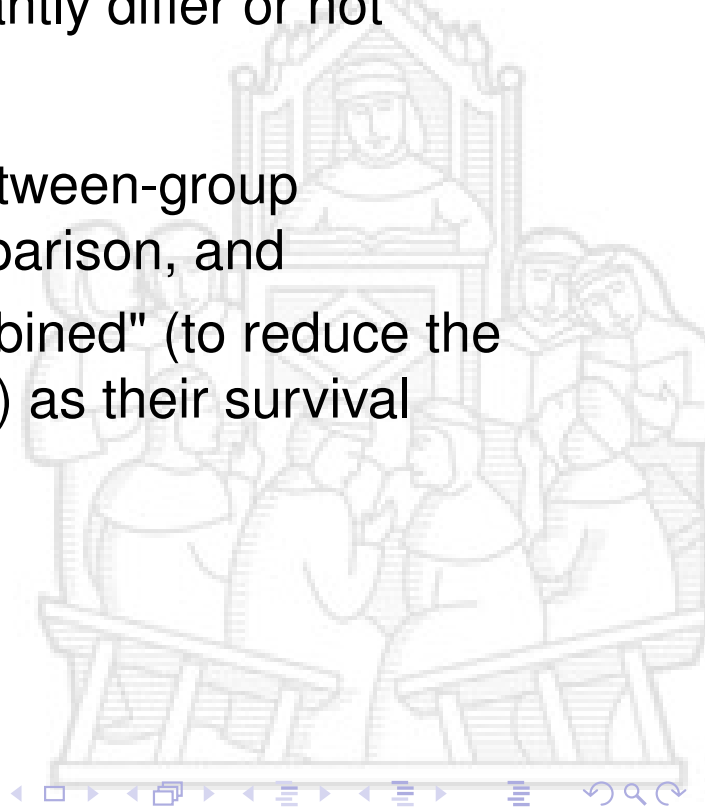
Under this setting the *G* statistic used to evaluate the significance of the overall model coefficients, when resulting statistically significant, suggests that at least one of the estimated coefficients has a hazard rate that is statistically significant from the reference group, not all are significant (!)

Cox proportional hazard model

When using a categorical variable, we may also be interested in understanding whether two of the $\hat{HR}_k$ significantly differ or not

This could be of interest

- to answer a specific question related to between-group comparison, in addition to the overall comparison, and
- to understand if two groups could be "combined" (to reduce the number of possible values of the covariate) as their survival experience does not differ



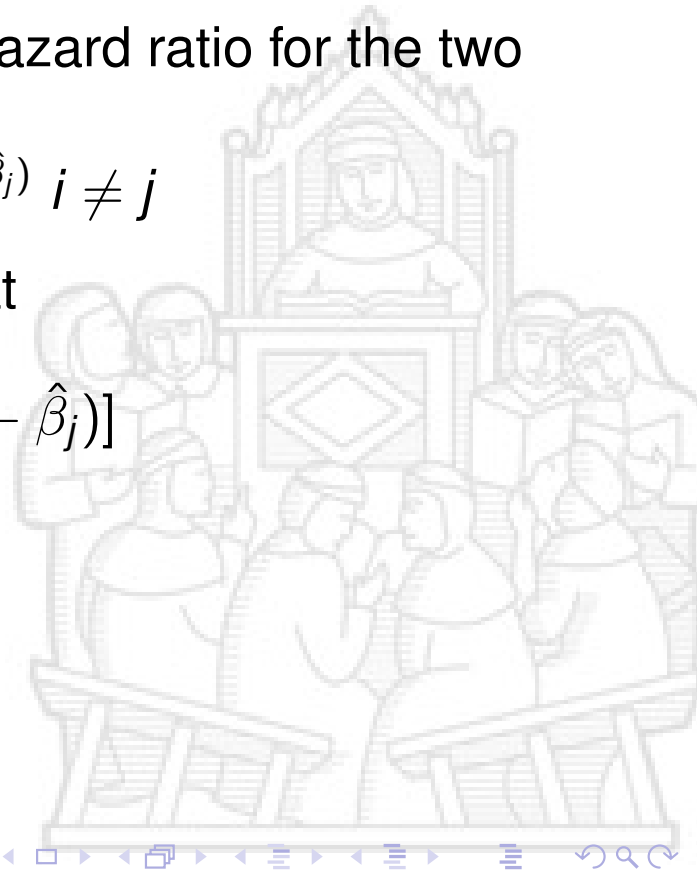
Cox proportional hazard model

This could be done using the estimator of the hazard ratio for the two groups

$$\hat{HR}(k = i \; versus \; k = j) = e^{(\hat{\beta}_i - \hat{\beta}_j)} \; i \neq j$$

and its 95% confidence interval considering that

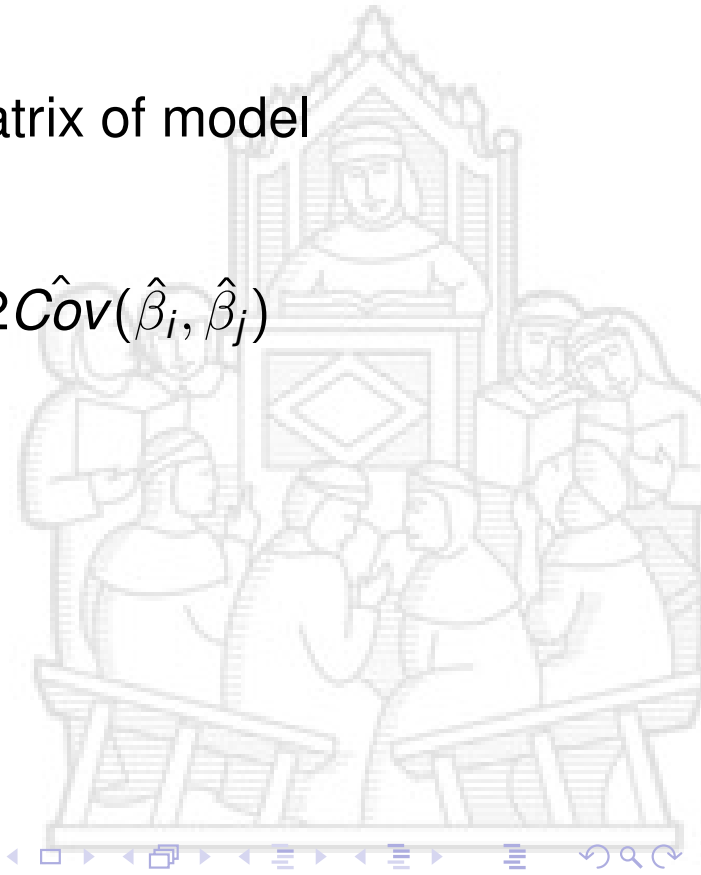
$$exp[(\hat{\beta}_i - \hat{\beta}_j) \pm 1.96 \times \hat{SE}(\hat{\beta}_i - \hat{\beta}_j)]$$



Cox proportional hazard model

Calculating $\hat{SE}(\hat{\beta}_i - \hat{\beta}_j)$ from the covariance matrix of model coefficients considering that

$$\hat{Var}(\hat{\beta}_i - \hat{\beta}_j) = \hat{Var}(\hat{\beta}_i) + \hat{Var}(\hat{\beta}_j) - 2\hat{Cov}(\hat{\beta}_i, \hat{\beta}_j)$$



Cox proportional hazard model

The Wald test for the difference $\hat{\beta}_i - \hat{\beta}_j$ is

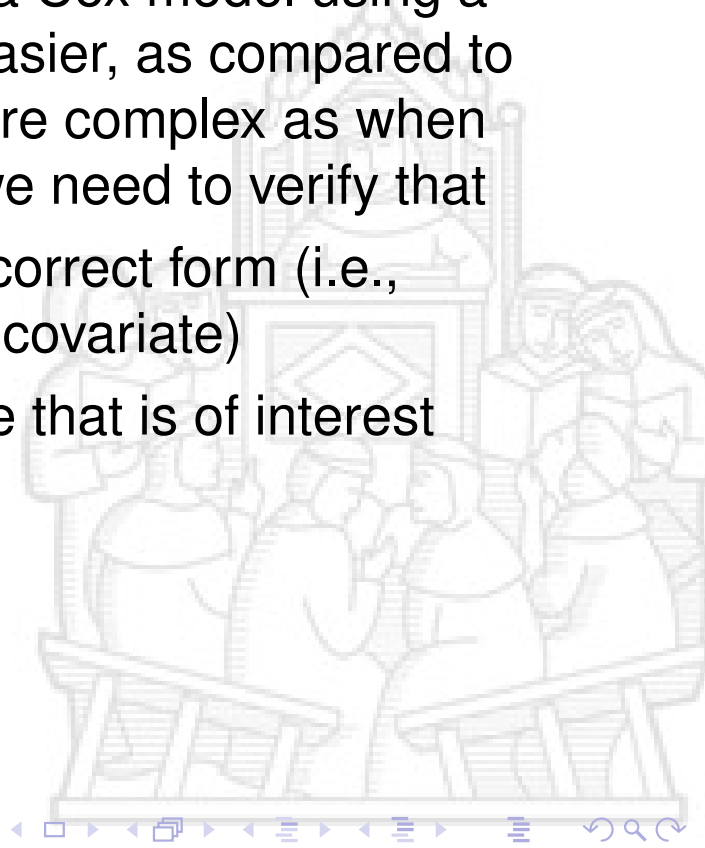
$$z = \frac{\hat{\beta}_i - \hat{\beta}_j}{\hat{SE}(\hat{\beta}_i - \hat{\beta}_j)}$$

which follows the standard normal distribution

Cox proportional hazard model

While the interpretation of model coefficient in a Cox model using a single countinuos covariate could be thought easier, as compared to dummy of categorical variables, it is indeed more complex as when handling continuos covariate in a Cox model, we need to verify that

- the variable is included in the model in its correct form (i.e., assume that the log hazard is linear in the covariate)
- the meaningful unit change in the covariate that is of interest (i.e., 1-unit change, 5-unit change)



Cox proportional hazard model

The inclusion of the variable in the model in its correct form needs to be properly verified, i.e., assessing the linear assumption

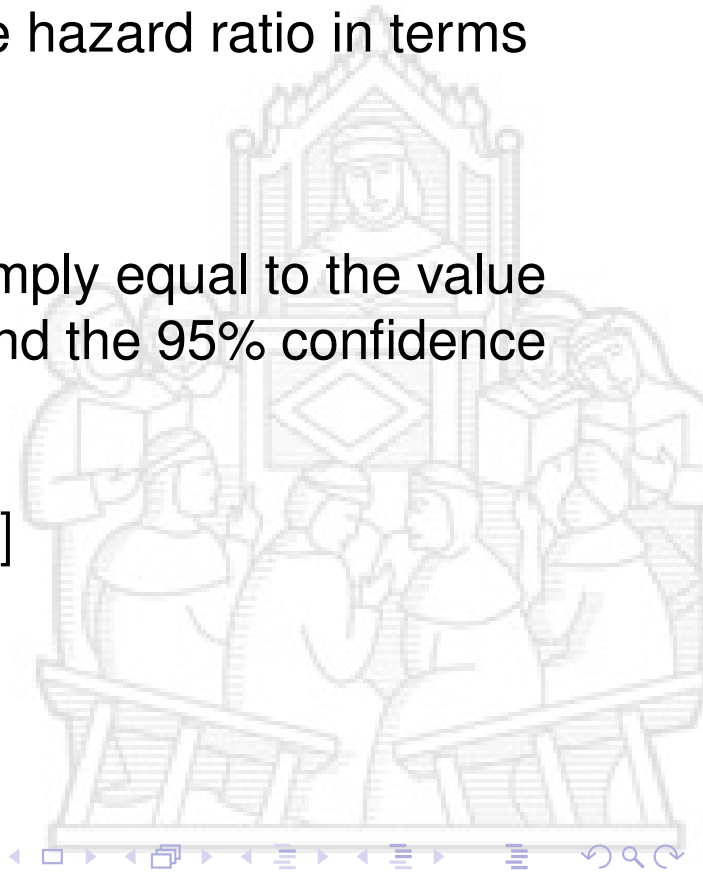
Cox proportional hazard model

While, if the meaningful unit change is in terms of cx (where c is a constant), we need to take into account that the hazard ratio in terms of cx is given by

$$\hat{HR}(c) = e^{c\hat{\beta}}$$

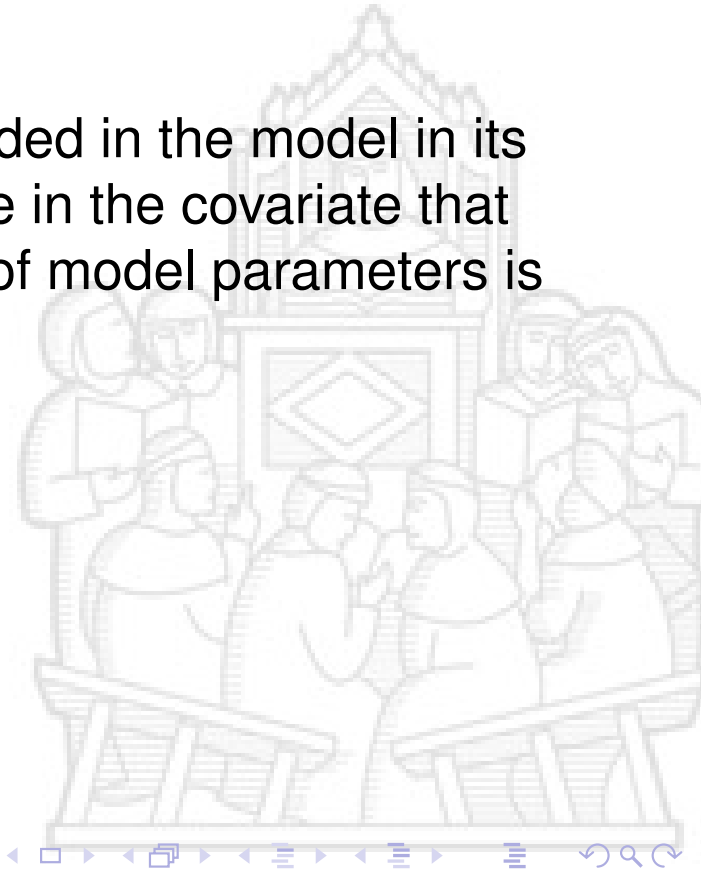
so the change in terms of cx rather than c is simply equal to the value of the change of interest times the coefficient and the 95% confidence interval could be obtained using

$$exp[c\hat{\beta} \pm z_{1-\alpha/2}|c|\hat{SE}(\hat{\beta})]$$



Cox proportional hazard model

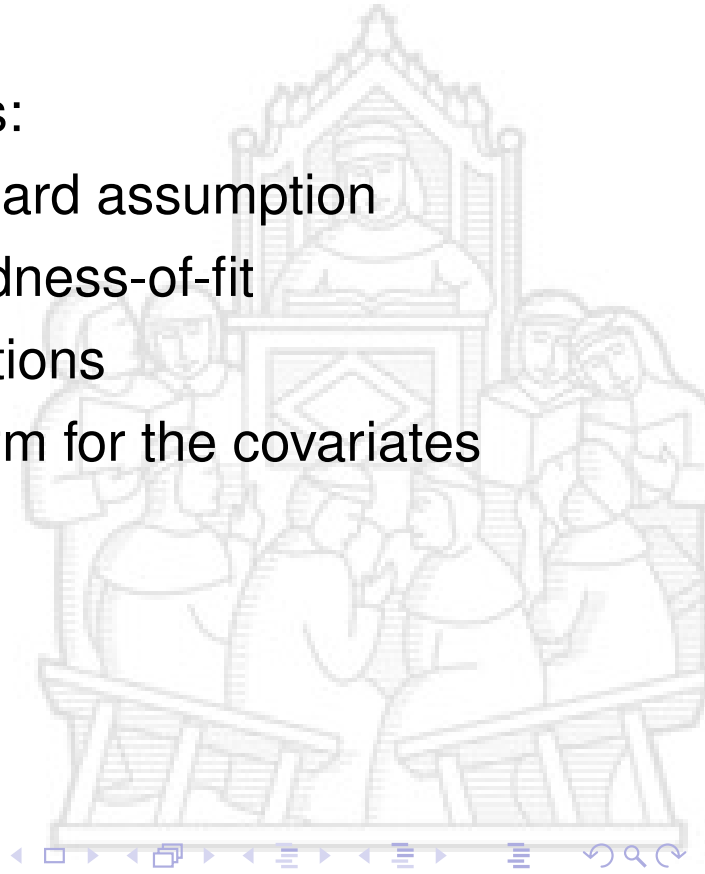
Only once we have verified the variable is included in the model in its correct form and we the meaningful unit change in the covariate that is of interest have been defined, interpretation of model parameters is straightforward



Assessing adequacy of a Cox proportional hazard model

Evaluating adequacy of the Cox model involves:

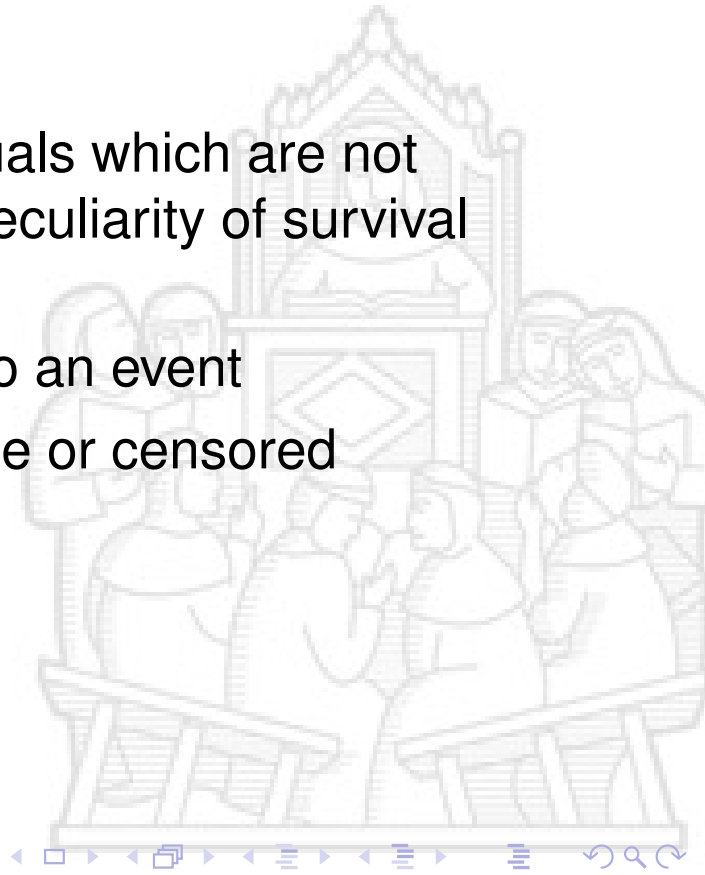
- assessing and testing the proportional hazard assumption
- computing summary measures of the goodness-of-fit
- assessing leverage and influence observations
- assessing the correctness of the model form for the covariates (i.e., linearity assumption)



Assessing adequacy of a Cox proportional hazard model

All these evaluation could be done using residuals which are not straightforward to be defined considering the peculiarity of survival analysis, because

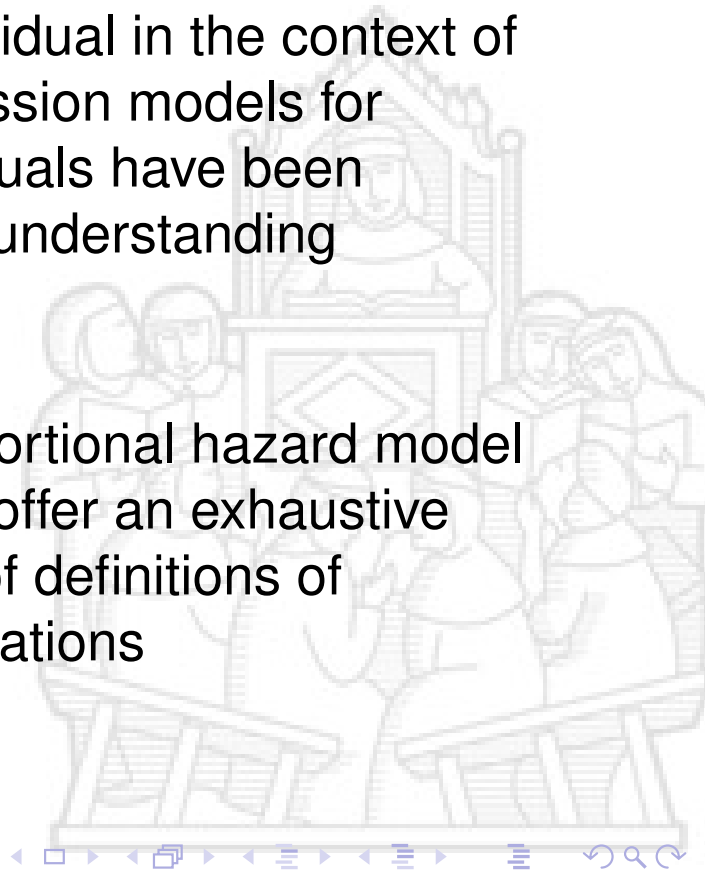
- the dependent variable of interest is time to an event
- the observed values are tipically incomplete or censored



Assessing adequacy of a Cox proportional hazard model

As there is no an obvious way to determine residual in the context of a Cox proportional hazard model (and in regression models for survival data in general), several different residuals have been roposed and each serves to a different goal in understanding adequacy of a fitted proportional hazard model

Some examples of residuals defined for a proportional hazard model are given in the next slides without claiming to offer an exhaustive overview, but just to illustrate some examples of definitions of residuals that may be useful for practical applications



Assessing adequacy of a Cox proportional hazard model

Schoenfeld and scaled Schoenfeld residuals are generally used to assess the proportional hazard assumption

Having observed a set of survival data which include some covariates x_i, the Shoenfeld residuals for the generic ith subject for the kth covariate is

$$\hat{r_{ik}} = c_i(x_{ik} - \hat{\bar{x}}_{w,k})$$

where c_i indicating censoring and being equal to zero for censored subjects, and

$$\hat{\bar{x}}_{w,k} = \frac{\sum x_{jk} e^{x_i' \hat{\beta}}}{\sum e^{x_i' \hat{\beta}}}$$

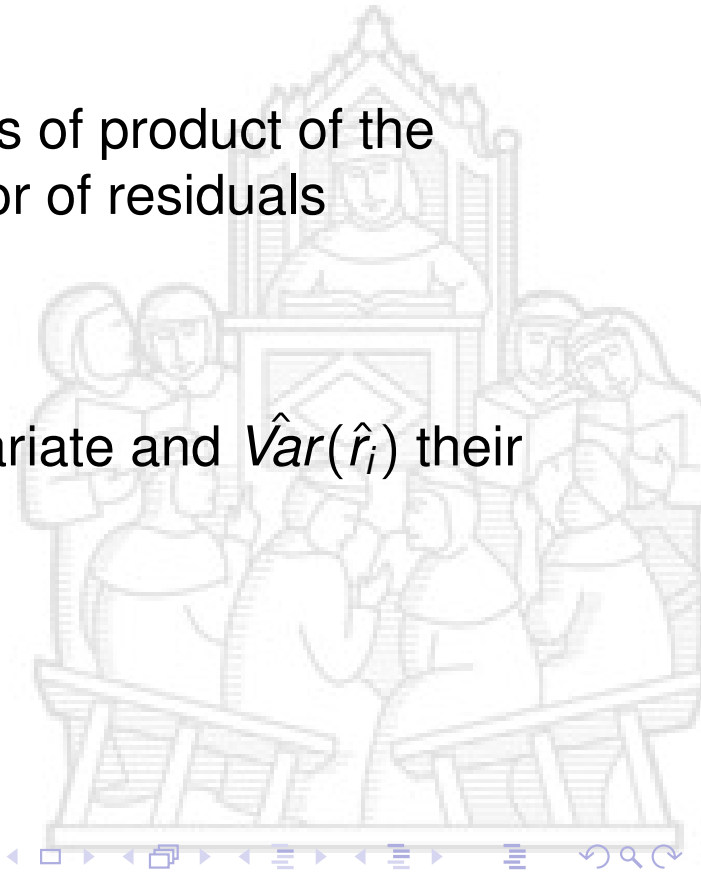
the sum of Schoenfeld residuals is zero and they are equal to zero for all censored subjects

Assessing adequacy of a Cox proportional hazard model

Scaled Shoenfeld residuals are defined in terms of product of the inverse of the covariance matrix times the vector of residuals

$$\hat{r}_i^* = [\hat{Var}(\hat{r}_i)]^{-1}\hat{r}_i$$

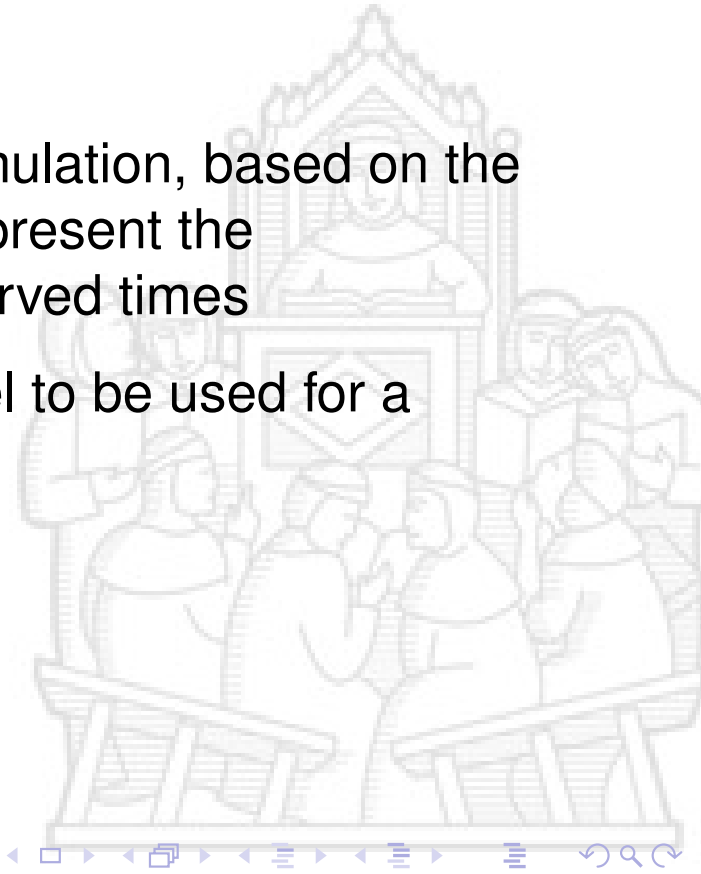
where $\hat{r}_i$ is the vector of residuals for the k covariate and $\hat{Var}(\hat{r}_i)$ their covariance matrix



Assessing adequacy of a Cox proportional hazard model

Martingale residuals have a more complex formulation, based on the possibility of having a counting process that represent the proportional hazard model at the different observed times

They are useful to understand the proper model to be used for a covariate included in the model



Developing a Cox proportional hazard model

In practical application, we are generally interested in including multiple covariates in a Cox model, this is done not only to assess the effect of multiple covariates of interest, but also in the case we are interested in a single covariate but we need to "control for confounding"

Developing a Cox proportional hazard model

Indeed, whn using multiple variable in a model (either Cox or whatever regression model), different situations may occur for a covariate x_j that could results

- a confounder of the relation between the main covariate of interest x_i and the outcome of interest
- an effect modifier of the relation
- nor a confounder, nor an effect modifier

Developing a Cox proportional hazard model

Broadly speaking, to understand the situation we are dealing with with respect to the covariates x_i and x_j, we could proceed in practical terms by fitting different regression models, including

- x_i but not x_j as covariate
- x_j but not x_i as covariate
- both x_i and x_j as covariates
- x_i,x_j and thier interaction $x_i \times x_j$ as covariates

Developing a Cox proportional hazard model

By assessing the significance (using the Wald statistic) of the interaction term $x_i \times x_j$ we are able to understand if there is an effect modification, if this could be excluded by comparing the change in model coefficient in models containg both covariates as compared to the one containing just one covariate we can understand if there is confounding

When more than one covariate is included in a model collinearity also needs to be asessed