**Survival Analysis**

subject had an event time = time to the event

subject did not have an event time = time followed in the study

linear regression cannot incorporate censored observations and distribution of survival time is highly skewed due to some people surviving an inordinate amount of time

logistic regression only considers whether an outcome occurred

survival analysis considers time until event occurred and can incorporate censored observations

**Censoring**

subject did not have an event during the period of time they were followed

Type I Censoring

observations are censored after a predetermined follow-up period

e.g. subjects who did not have the event of interest within 2 years are censored

Type II Censoring

observations are censored after a fixed percentage of subjects develop the event of interest

e.g. study will keep monitoring subjects until 10% of people have the event of interest, then censor all remaining subjects

Random Censoring

observations are censored for reasons outside the control of investigators

censoring that’s not part of the study design

e.g. subject moved out of the country

**Survival Analysis Example**

study follows subjects for up to 2 years with death as the event of interest

Subject 1 alive at the end of study T = 24 censored observation

Subject 2 dropped out after a year T = 12 censored observation

Subject 3 died after 10 months T = 10 observed event

Subject 4 died after 21 months T = 21 observed event

Subject 1 survived at least 24 months and Subject 2 survived at least 12 months

Subject 4 survived 11 months longer than Subject 3 before dying

**Random Censoring**

Informative Censoring

people who are censored would have had different outcomes than those who remained in the analysis for the same amount of time

censoring due to competing risks is usually informative because they’re related to the reason for leaving the study

e.g. subject dropped out after 6 months because they’re too sick to continue study visits, so probably had a higher risk of death than similar people who remained in the study for at least 6 months

Non-Informative Censoring

people who are censored would have similar risk for the outcome as those who remained in the analysis for the same amount of time

e.g. subject moved out of the country after 6 months and researchers have no reason to believe that person would have a different risk for disease than similar people who remained in the study for at least 6 months

basic survival analysis assumes that censoring is non-informative

**Survival and Hazard Functions**

T = survival time to event

survival distribution (subject survives at least to time )

hazard function = instantaneous rate of occurrence of the event

= density of time to event

cumulative hazard

**Nonparametric Approach**

no assumptions are made on the shape of the underlying distribution for survival time

Kaplan-Meier Curves/Product-Limit Estimate for descriptive analysis

log-rank test for crude analytical comparison among several groups

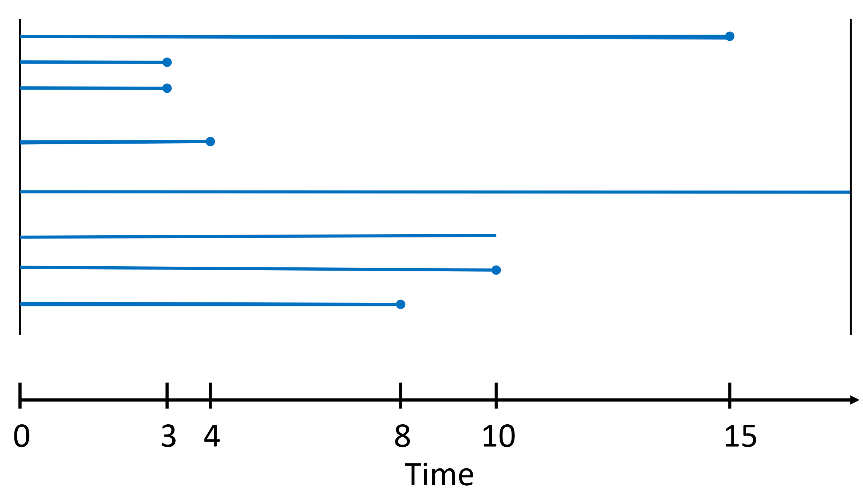
**Kaplan-Meier Estimation**

crude comparison between two groups

doesn’t provide an effect estimate or adjust for covariates

partition time axis according to when events occur

e.g. event times are 3, 4, 8, 10, and 15



|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Time Interval** | **# Fail** | **# Survive** | **# Censored** | **# Remain** |
| 0 | 0 | 100 | 0 | 100 |
| 1 | 5 | 95 | 5 | 90 |
| 2 | 10 | 80 | 0 | 80 |
| 3 | 12 | 68 | 3 | 65 |

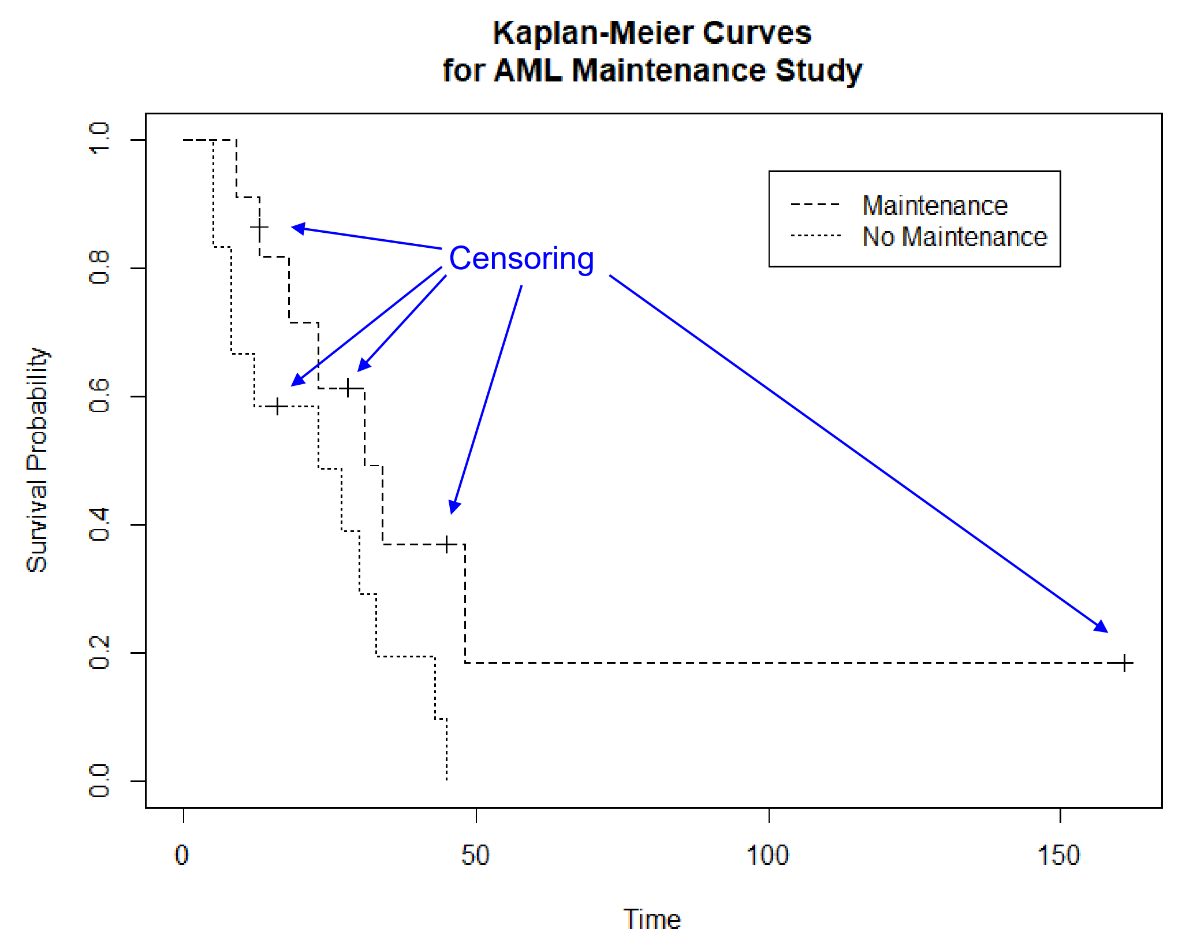
T = 0 start with 100

T = 1 start with 100, but 5 died and 5 were censored so 90 left

T = 2 start with 90, but 10 died so 80 left

T = 3 start with 80, but 12 died and 3 were censored, so 65 left

calculate survival function using product of conditional probabilities



Summary Measures

median survival time = time where

median survival time can’t be calculated if less than half the subjects have the event

mean survival is often biased because survival time for all subjects are not calculated

hazard ratio cannot be estimated from the Kaplan-Meier curve and depends on the proportional hazards assumption

**Log-Rank Test**

non-parametric test that compares the survival distributions in 2 or more groups

time-stratified Mantel Extension chi-square test

compares observed events with expected number of events under the null hypothesis of no difference in survival between the two groups

doesn’t measure association between groups

H0: The survival distribution for both groups are the same

The hazard functions for both groups are the same.

H1: The survival distribution for one group is a power of the other.

The hazard function for one group is a multiple of the other group’s hazard function.

|  |  |  |  |
| --- | --- | --- | --- |
| **Failure Time** | | | |
| **Group** | **Observed events at** | **Surviving Beyond** | **At Risk at** |
| 1 |  |  |  |
| 2 |  |  |  |
| Total |  |  |  |

expected events in Group 1 expected events in Group 2

total observed events in Group 1

total expected events in Group 1

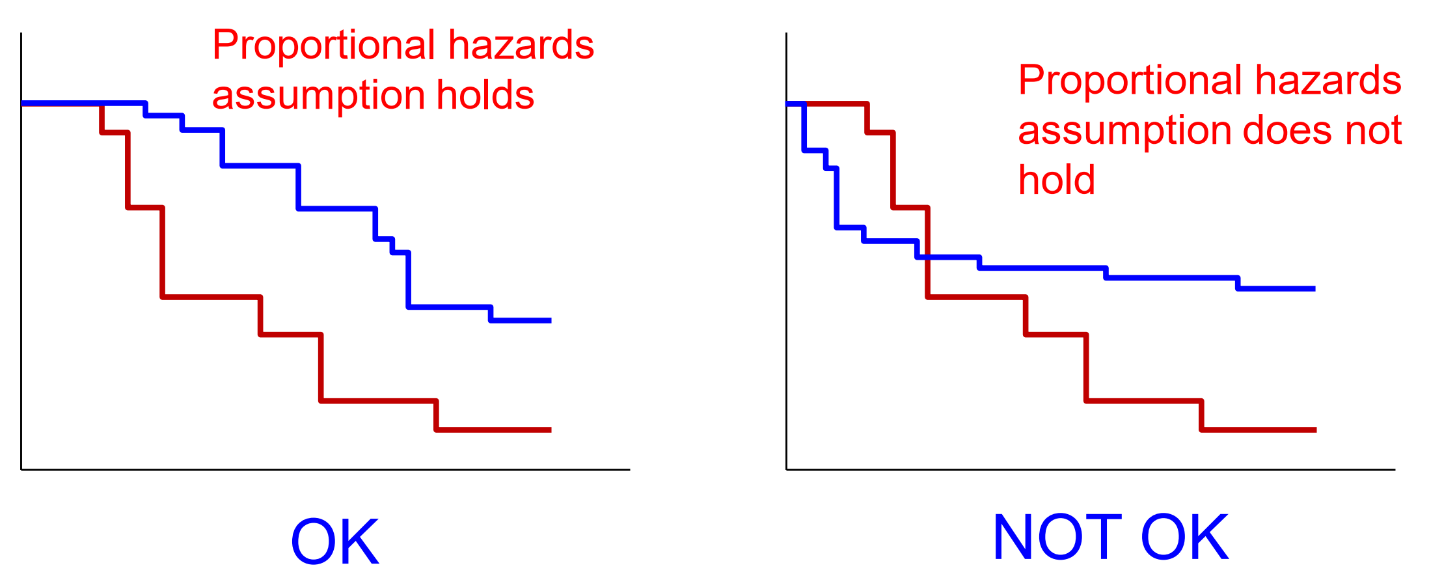
Log-Rank statistic =

= #groups – 1

**Proportional Hazards Assumption**

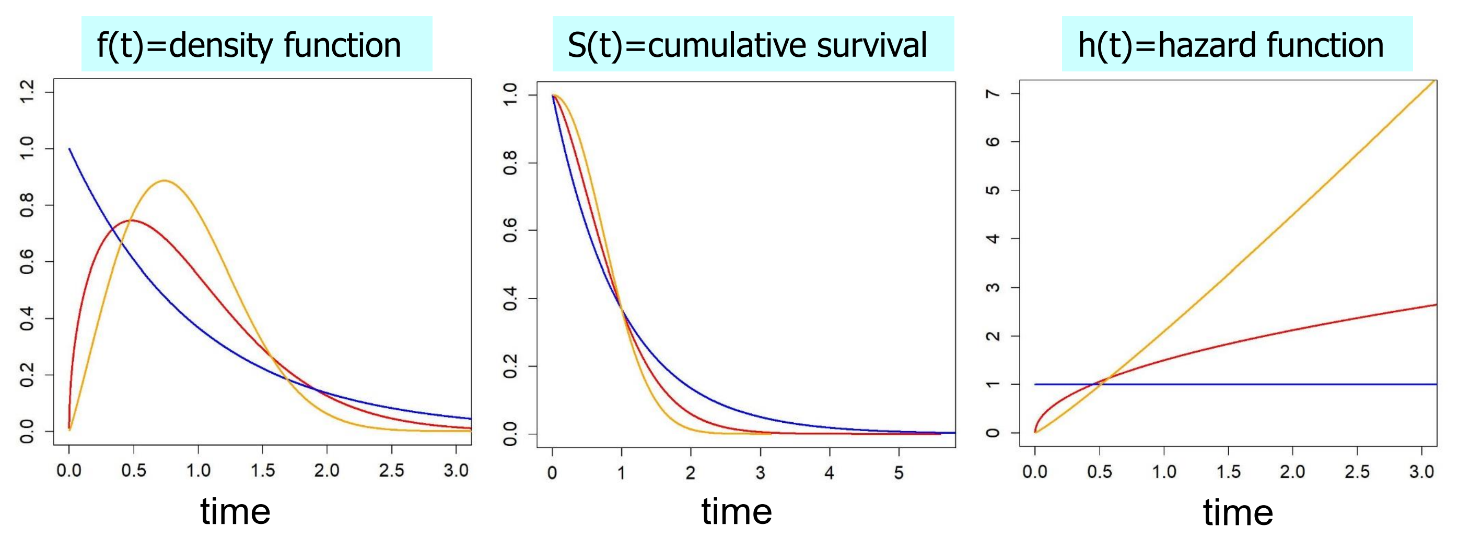
hazard functions in different groups are proportional

survival distributions crossing is an indication of non-proportional hazards



**Weibull Model**

Weibull distribution is very flexible and can take many shapes

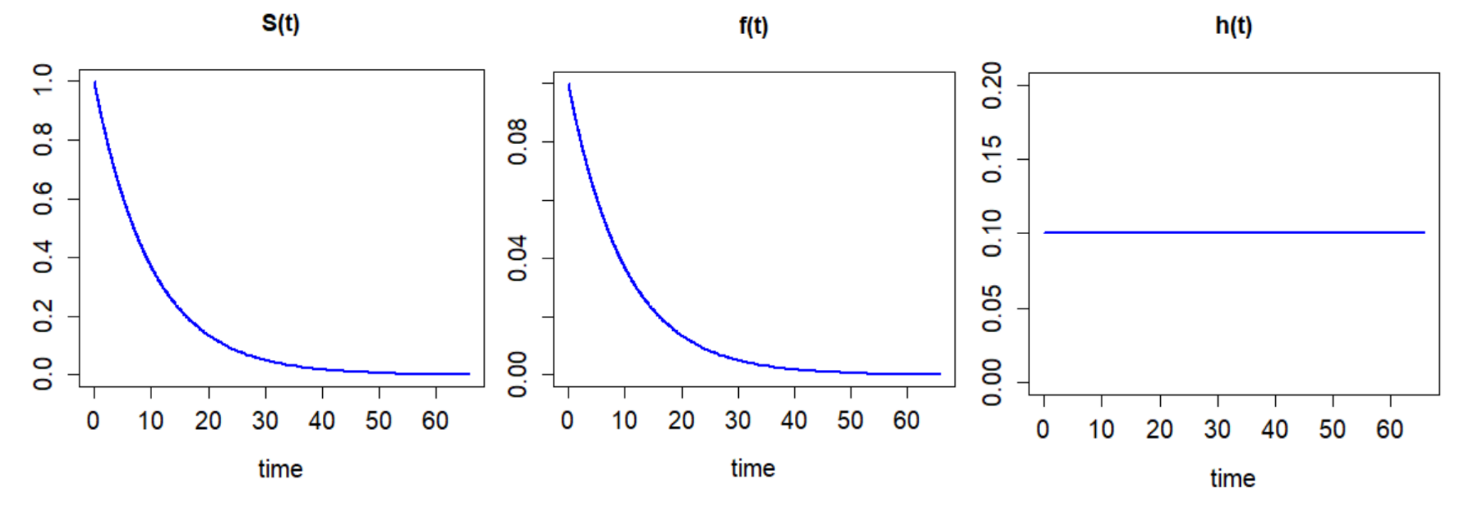


Exponential Model

simplest parametric model

hazard function doesn’t depend on time so is constant

proportional hazards model because hazards ratio is the same at all times



survival function

density function

hazard function

model the hazard as a function of the exposure to quantify the relative hazard

hazard for disease in unexposed

hazard for disease in exposed

hazard ratio

probability of surviving free of the disease until age t in unexposed

probability of surviving free of the disease until age t in exposed

model assumes hazard of the disease at age t, given no exposure before age t, is constant

not reasonable because the older someone is the more likely he is to develop the disease, so hazard of exposure should increase with time

**Proportional Hazards Models**

Exponential Model

= baseline hazard × effects of covariates

baseline hazard is constant because it doesn’t change with time

General Models

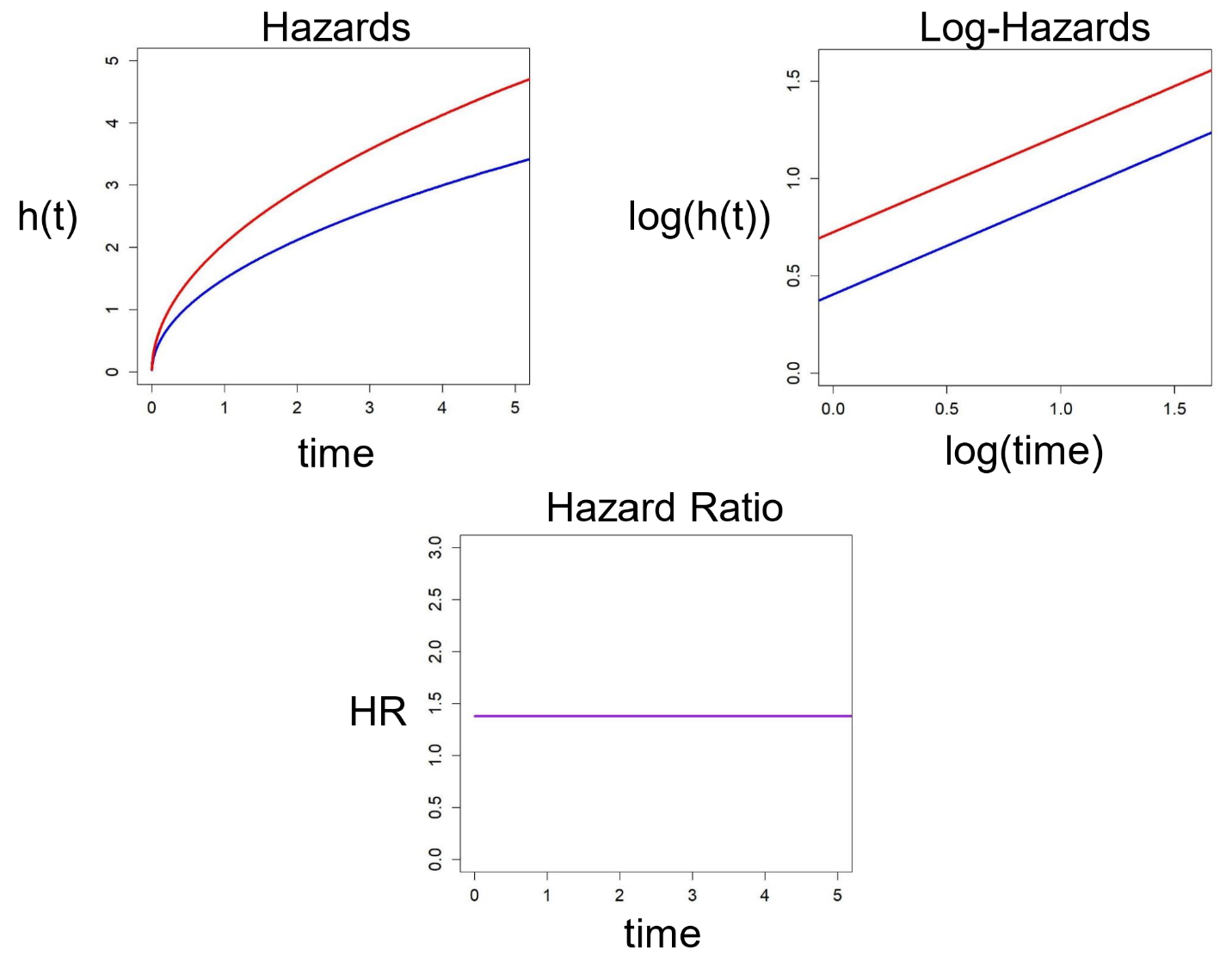
= baseline hazard × effects of covariates

baseline hazard is a function of time

hazard ratio at time t for a change in X

logarithmic curves are parallel and separated by

hazard ratio doesn’t depend on time



**Cox Proportional Hazards Model**

baseline hazard function is treated as a nuisance function and is left uncalculated

covariates affect the hazard function multiplicatively through the function

semi-parametric model baseline hazard function non-parametric

effects of covariates parametric

suitable when parameter estimates of the covariates are more important than the shape of the hazard

fit by maximizing the partial likelihood function

Single Variable

baseline hazard for unexposed subjects

baseline hazard for exposed subjects

hazard ratio of exposed vs unexposed

hazard ratio comparing two specific values of

95% confidence interval

Multiple Covariates

time is in the baseline hazard

covariates are in the exponentiated multiplier of the baseline hazard

shape of the baseline hazard is undefined

subject with values of 0 for every covariate

all subjects’ hazard relative to the baseline hazard

**Testing Proportional Hazards Assumption**

Graphical Assessment

curves in plot of natural log of vs time are parallel for all values of X and separated by

curves in plot of vs natural log of time are parallel for all values of X and separated by

Schoenfeld Residuals

H0: The proportional hazard assumption is satisfied

H1: The proportional hazard assumption is not satisfied.

Schoenfeld residual for subject who had an event at time

estimated mean of X based on the subjects at risk at time

scaled Schoenfeld residuals should be uncorrelated with time

curve of covariate vs time should be approximately a horizontal line

Time-Varying Effects

H0: is not a linear function of time. The proportional hazard assumption is satisfied

H1: is a linear function of time. The proportional hazard assumption is not satisfied.

fit a model with time-varying effects, allowing coefficients to change with time, approximated by time-varying covariates

time-varying effect effect of a variable changes as a function of time

time-varying covariate variable changes as a function of time

model includes an interaction term between time and variable

if interaction term is not 0, then the corresponding term fails the proportional hazard function and is a linear function of time

**Accounting for Non-Proportional Hazards**

Stratified Analysis

variable failing proportional hazards assumption aren’t of interest

each stratum has the same but different baseline hazard

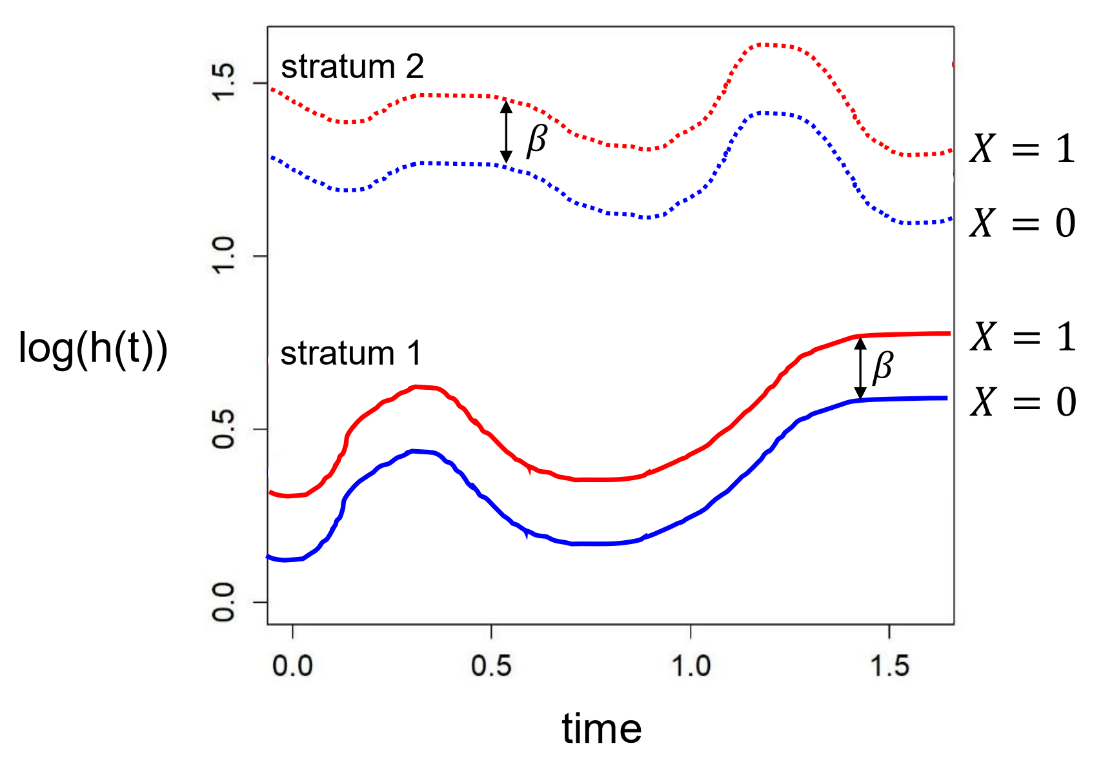
effect of the variable is now part of the baseline hazard, which is not estimated

cannot use the model to calculate a hazard ratio between subjects in different strata

if the variable is continuous, stratify it using a categorical variable and include the continuous variable as a covariate in the model to account for residual association within the strata

Stratum 1

Stratum 2

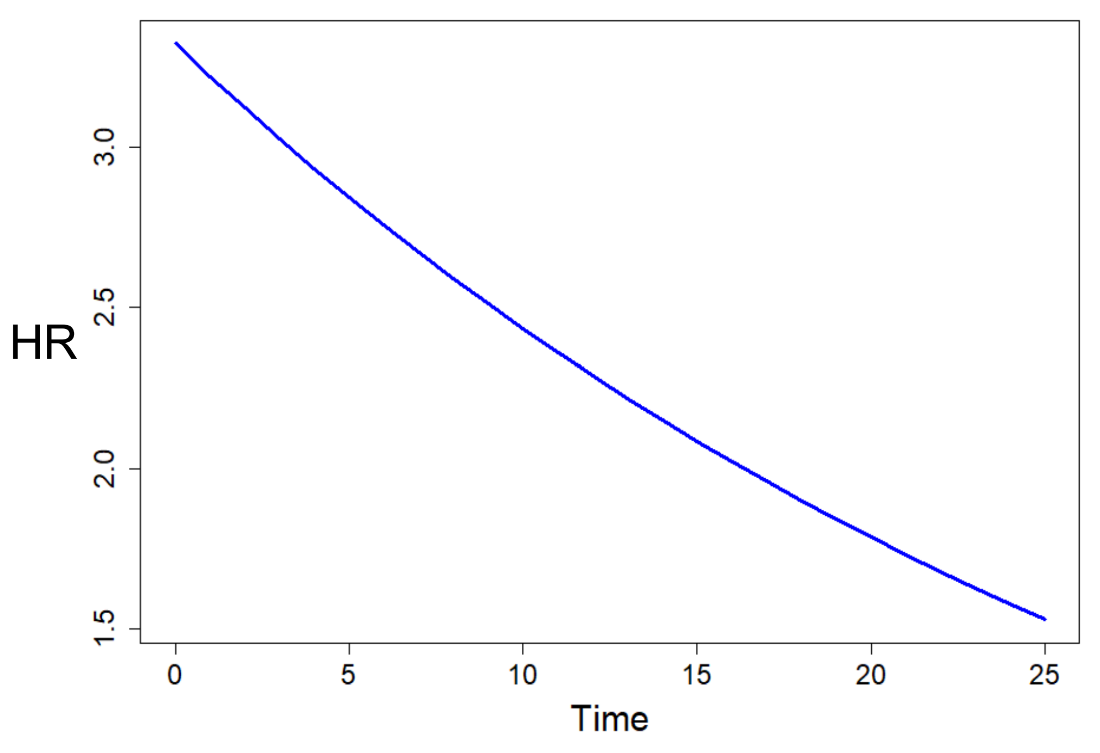


The hazard ratio for the disease when comparing exposed subjects to unexposed subjects within the same stratum is.

Time-Dependent Variable

variables failing the proportional hazards assumption are of interest

include interaction term between variable and time in model



hazard ratio comparing exposed to unexposed changes over time

over time, the effect of the variable on the outcome decreases