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 $S(t) = Pr(T > t) = Pr(\text{subject survives at least to time } t)$

hazard function = instantaneous rate of occurrence of the event

$$h(t) = \lim_{\Delta t \to 0} \frac{Pr(t < T \leq t + \Delta t | T > t)}{\Delta t}$$

$f(t)$ = density of time to event

$$h(t) = \frac{f(t)}{S(t)}$$

cumulative hazard $H(t) = -ln\big(S(t)\big)$

Nonparametric Approach

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

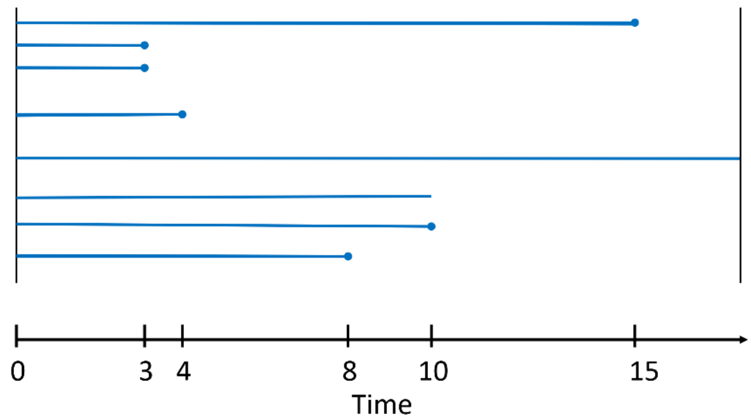
Kaplan-Meier Curves/Product-Limit Estimate for descriptive analysis

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

$$\#remain = \#survive - \#censored$$

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

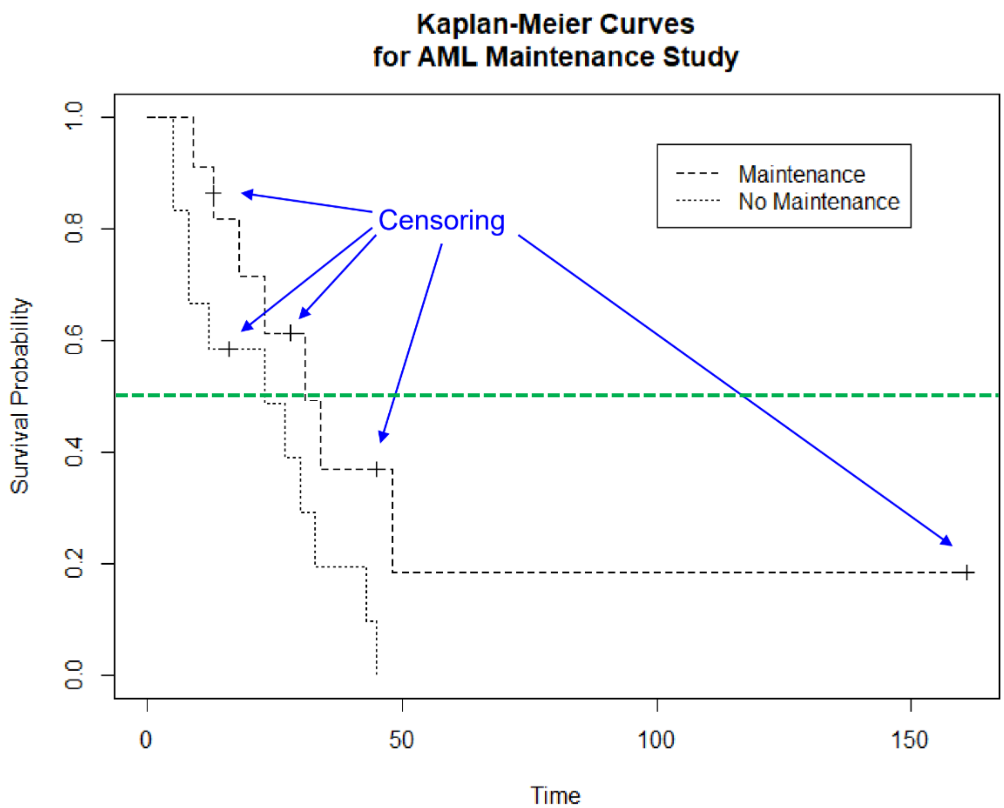
calculate survival function using product of conditional probabilities

$$S(0) = Pr(T > 0) = \frac{100}{100} = 1.00$$

$$S(1) = Pr(T > 1) = \frac{95}{100} = 0.95$$

$$S(2) = Pr(T > 2) = Pr(T > 1) \times Pr(T > 2|T > 1) = \left(\frac{95}{100}\right)\left(\frac{80}{90}\right) = 0.8444$$

$$S(3) = Pr(T > 3) = Pr(T > 1) \times Pr(T > 2|T > 1) \times Pr(T > 3|T > 2) = \left(\frac{95}{100}\right)\left(\frac{80}{90}\right)\left(\frac{68}{80}\right)$$
$$= 0.7178$$



<u>Summary Measures</u>
median survival time = time where $S(t) < 0.5$
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

H_0: $S_1(t) = S_2(t)$	The survival distribution for both groups are the same
$h_1(t) = h_2(t)$	The hazard functions for both groups are the same.
H_1: $S_1(t) = \left(S_2(t)\right)^{\theta}$	The survival distribution for one group is a power of the other.
$h_1(t) = \theta h_2(t)$	The hazard function for one group is a multiple of the other group's hazard function.

j^{th} Failure Time			
Group	**Observed events at t_j**	**Surviving Beyond t_j**	**At Risk at t_j**
1	o_{1j}	$n_{1j} - o_{1j}$	n_{1j}
2	o_{2j}	$n_{2j} - o_{2j}$	n_{2j}
Total	o_j	$n_j - o_j$	n_j

expected events in Group 1 $(e_{1j}) = \frac{o_j n_{1j}}{n_j}$ expected events in Group 2 $(e_{2j}) = \frac{o_j n_{2j}}{n_j}$

$$var(o_{1j}) = \frac{n_{1j} n_{2j} o_j (n_j - o_j)}{n_j n_j (n_j - 1)}$$

total observed events in Group 1 $(O_1) = \sum_j o_{1j}$

total expected events in Group 1 $(E_1) = \sum_j e_{1j}$

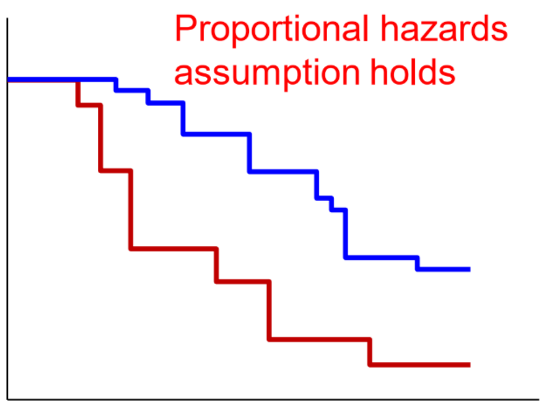
$$var(o_{1j}) = \sum_j v_{1j}$$

Log-Rank χ^2 statistic = $\frac{(O_1 - E_1)^2}{V}$

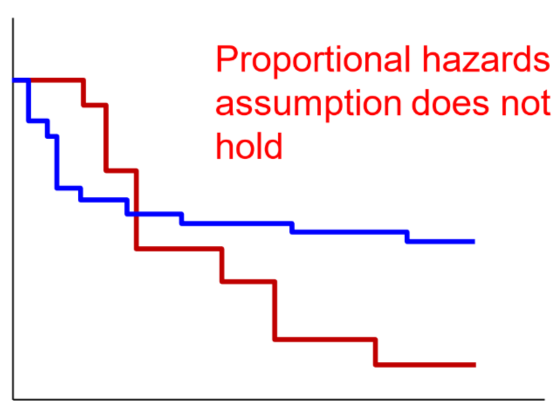
df = #groups – 1

Proportional Hazards Assumption

hazard functions in different groups are proportional
survival distributions crossing is an indication of non-proportional hazards



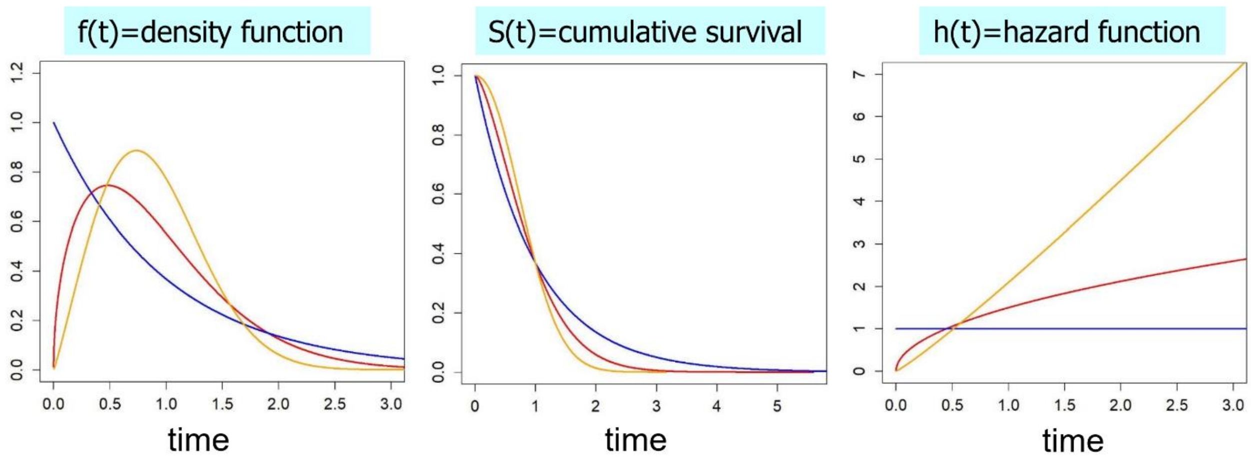
OK



NOT OK

Weibull Model

Weibull distribution is very flexible and can take many shapes

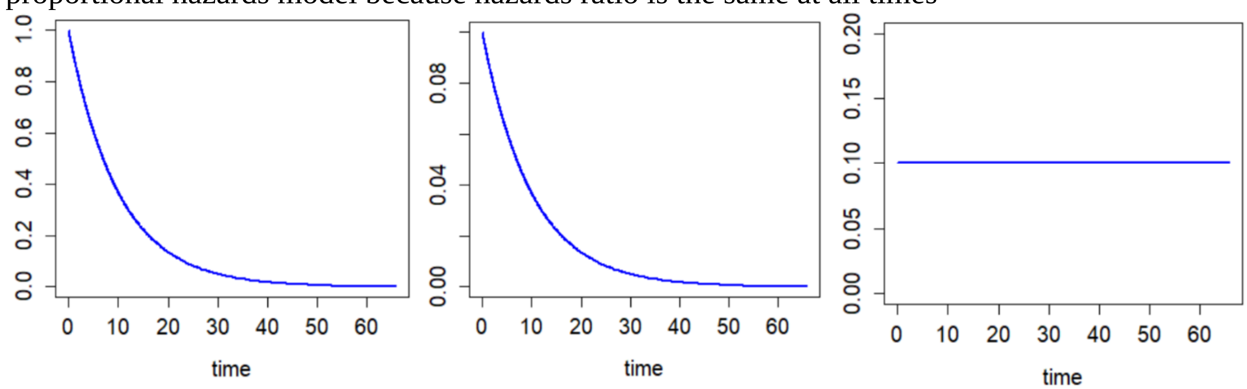


Exponential Model

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	$S(t\|\lambda) = P(T > t\|\lambda) = e^{-\lambda t}$
density function	$f(t\|\lambda) = \lambda e^{-\lambda t}$
hazard function	$h(t\|\lambda) = \lambda$

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

$$log\big(h(t|X)\big) = log(\lambda) = \beta_0 + \beta_1 X$$

e^{β_0}	hazard for disease in unexposed
$e^{\beta_0+\beta_1}$	hazard for disease in exposed
e^{β_1}	hazard ratio

$e^{\beta_0 t}$	probability of surviving free of the disease until age t in unexposed
$e^{(\beta_0+\beta_1)t}$	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

<u>Exponential Model</u>

$h(t|X) = e^{\beta_0} e^{\beta_1 X} = h_0 e^{\beta_1 X}$ = baseline hazard × effects of covariates

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

<u>General Models</u>

Model: $h(t|X) = \underbrace{h_0(t)}_{\text{Function of time}} \underbrace{e^{\beta_1 X}}_{\text{Function of X}}$ = baseline hazard × effects of covariates

baseline hazard is a function of time

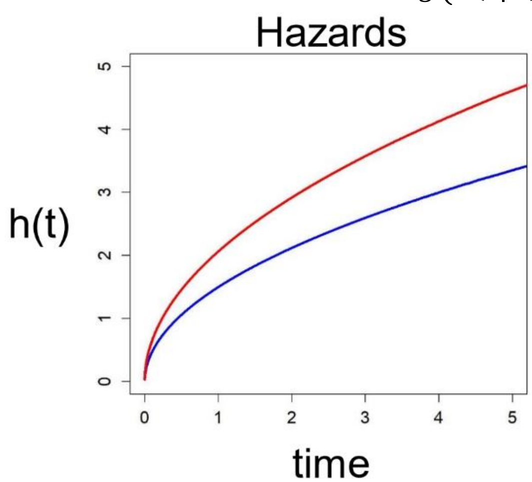
$$HR = \frac{h(t|X = x_1)}{h(t|X = x_2)} = \frac{h_0(t) e^{\beta_1 x_1}}{h_0(t) e^{\beta_1 x_2}} = e^{\beta_1 (x_1 - x_2)}$$

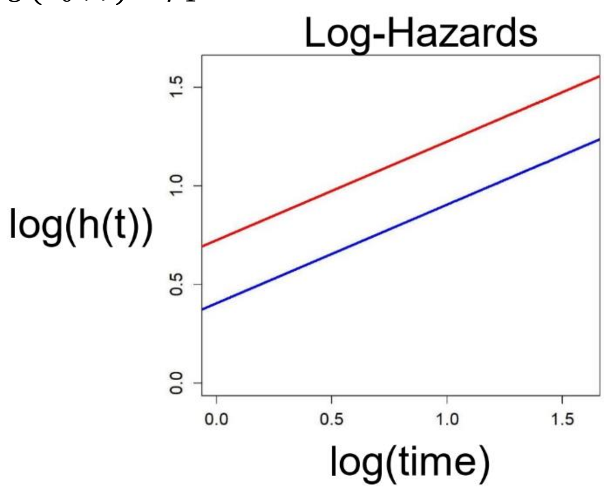
hazard ratio at time t for a change in X

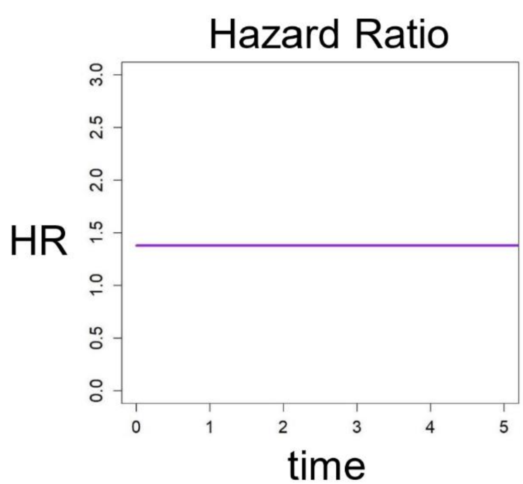
logarithmic curves are parallel and separated by β_1

hazard ratio doesn't depend on time

$$log\big(h(t|X)\big) = log\big(h_0(t)\big) + \beta_1 X$$







Cox Proportional Hazards Model

$$h(t|X) = h_0(t)e^{\beta X}$$

baseline hazard function $h_0(t)$ is treated as a nuisance function and is left uncalculated
covariates affect the hazard function multiplicatively through the function $e^{\beta X}$

semi-parametric model	baseline hazard function $h_0(t)$	non-parametric
	effects of covariates $e^{\beta X}$	parametric

suitable when parameter estimates of the covariates are more important than the shape of the hazard
fit by maximizing the partial likelihood function

<u>Single Variable</u>

$$h(t|X) = h_0(t)e^{\beta_1 X}$$

$h(t\|X = 0) = h_0(t)$	baseline hazard for unexposed subjects
$h(t\|X = 1) = h_0(t)e^{\beta_1 X}$	baseline hazard for exposed subjects
e^{β_1}	hazard ratio of exposed vs unexposed
$e^{\beta_1(x_1 - x_2)}$	hazard ratio comparing two specific values of X
$e^{(\beta_1 + 1.96SE(\beta_1))(x_1 - x_2)}$	95% confidence interval

<u>Multiple Covariates</u>

$$h(t|X_1, X_2, X_3 \ldots X_k) = h_0(t)e^{\beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \ldots \beta_k X_k}$$

time is in the baseline hazard $h_0(t)$
covariates are in the exponentiated multiplier of the baseline hazard $h_0(t)$
shape of the baseline hazard is undefined

$h_0(t)$	subject with values of 0 for every covariate
$e^{\beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \ldots \beta_k X_k}$	all subjects' hazard relative to the baseline hazard

Testing Proportional Hazards Assumption

<u>Graphical Assessment</u>

$$log\big(h(t|X)\big) = log\big(h_0(t)\big) + \beta X$$

curves in plot of natural log of $h(t|X)$ vs time are parallel for all values of X and separated by β

$$log\left(-log\big(S(t|X)\big)\right) = log\left(-log\big(S_0(t)\big)\right) + \beta X$$

curves in plot of $log\left(-log\big(S(t|X)\big)\right)$ vs natural log of time are parallel for all values of X and separated by β

<u>Schoenfeld Residuals</u>

H_0: The proportional hazard assumption is satisfied
H_1: The proportional hazard assumption is not satisfied.

$x_1 - \bar{x}(t_i)$	Schoenfeld residual for subject i who had an event at time t_i
$\bar{x}(t_i)$	estimated mean of X based on the subjects at risk at time t_i

scaled Schoenfeld residuals should be uncorrelated with time
curve of covariate vs time should be approximately a horizontal line

Time-Varying Effects

H_0: β is not a linear function of time. The proportional hazard assumption is satisfied

H_1: β 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	$\beta(t)X$	effect of a variable changes as a function of time
time-varying covariate	$\beta X(t)$	variable changes as a function of time

$$\beta = a + bt$$
$$h(t|X) = h_0(t)e^{\beta X} = h_0(t)e^{(a+bt)X} = h_0(t)e^{aX+btX}$$

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 β 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 β but different baseline hazard

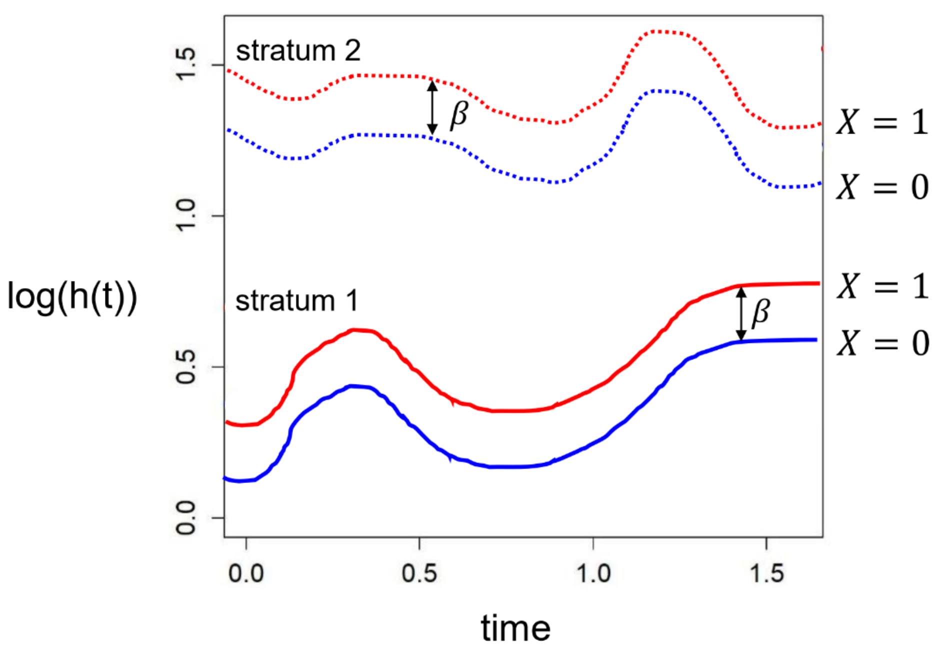
effect of the variable is now part of the baseline hazard, which is not estimated

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 $\quad h_1(t|X) = h_{01}(t)e^{\beta X}$

Stratum 2 $\quad h_2(t|X) = h_{02}(t)e^{\beta X}$



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

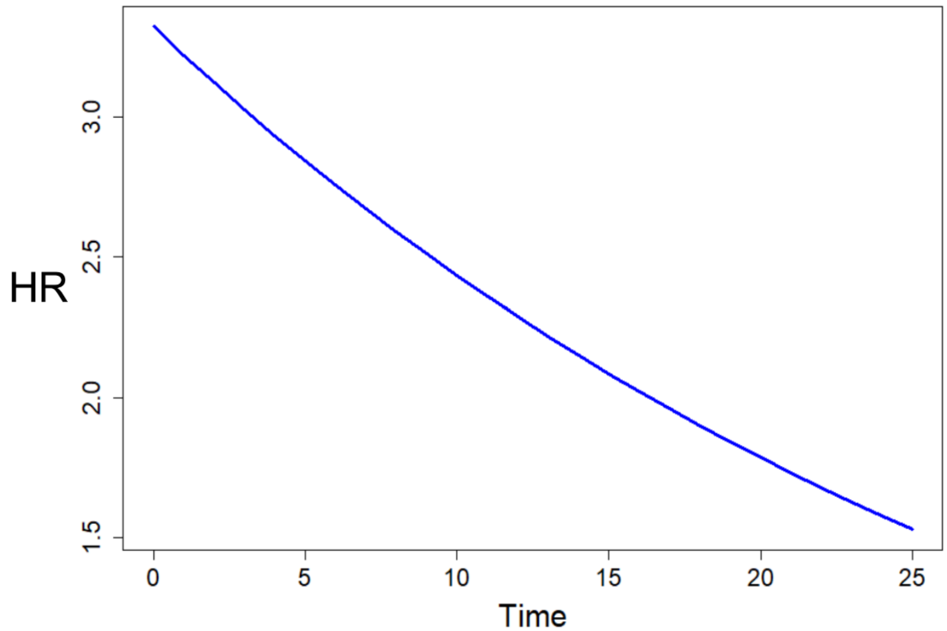
<u>Time-Dependent Variable</u>

variables failing the proportional hazards assumption are of interest

include interaction term between variable and time in model

$$h(t|X) = h_0(t)e^{aX+btX}$$

$$HR(t) = e^{(a+bt)}$$



hazard ratio comparing exposed to unexposed changes over time

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