

Survival analysis

Recommended texts

- **Survival analysis: A Self-learning text**
 - David G. Kleinbaum, Springer.
 - An excellent tutorial on survival analysis, including Cox proportional hazards
- **Primer of Biostatistics**
 - Stanton Glantz, McGraw-Hill
 - Chapter 11 gives a very understandable introduction to survival analysis

We use survival analysis:

- To describe the survival times of members of a group
 - Survival function
 - Hazard function
 - Kaplan-Meier curves

We use survival analysis:

- To compare the survival times of two or more groups
 - Log-rank test

We use survival analysis:

- To describe the effect on survival times of a continuous variable (e.g., gene expression)
 - Cox proportional hazard regression
 - Gives relative risk for a unit change in the variable
- E.g., a unit change in the expression of gene A gives a 2-fold increase in relative risk

Survival analysis: definitions

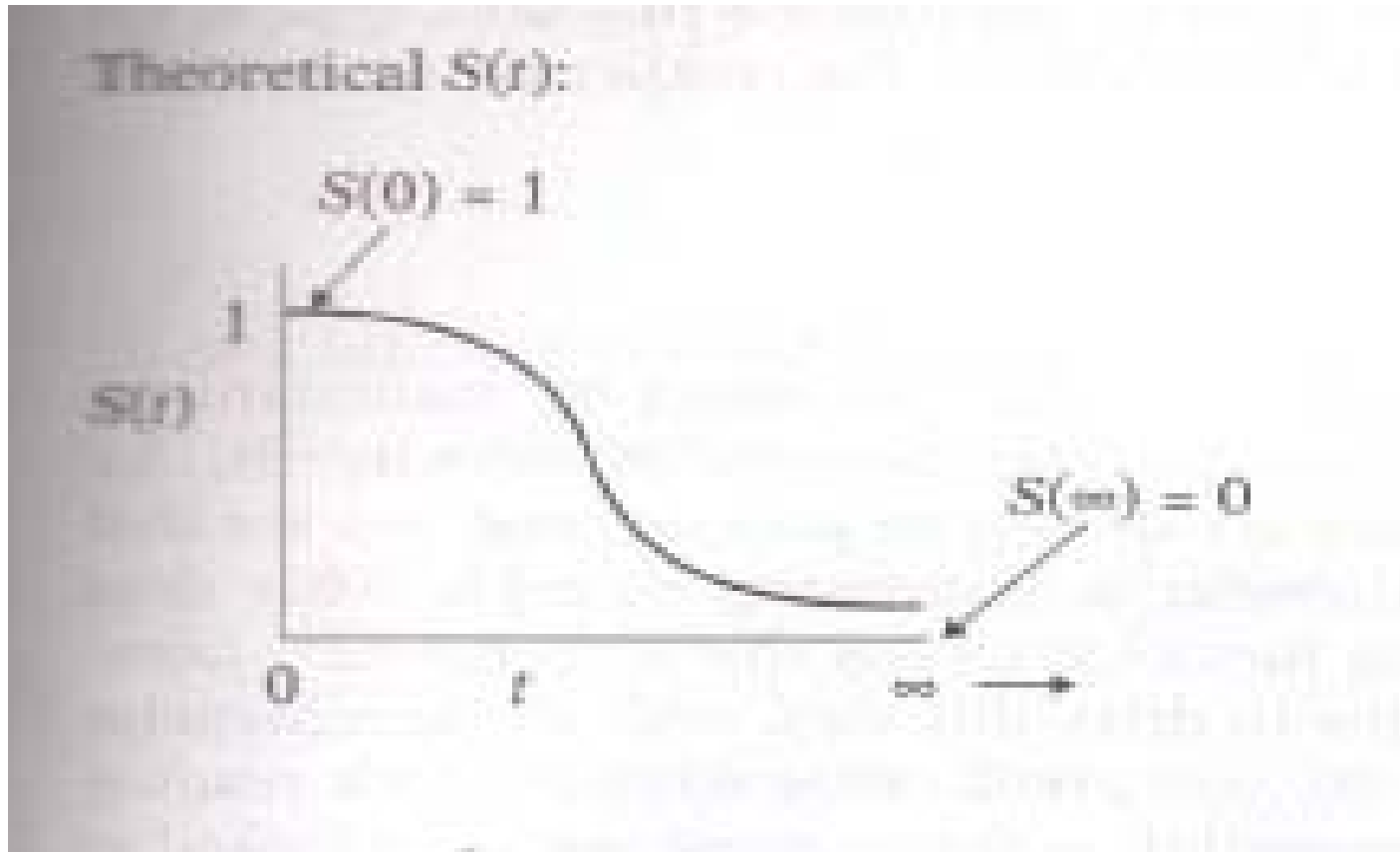
- Event: Death, disease occurrence, disease recurrence, recovery, or other experience of interest
- Time: The time from the beginning of an observation period (e.g., surgery) to (a) an event, or (b) end of the study, or (c) loss of contact or withdrawal from the study.

- Censoring / Censored observation: When a subject does not have an event during the observation time, they are described as censored, meaning that we cannot observe what has happened to them subsequently.
- A censored subject may or may not have an event after the end of observation time.

- Survivor function $S(t)$:
- $S(t)$ = The probability that a subject survives longer than time t .

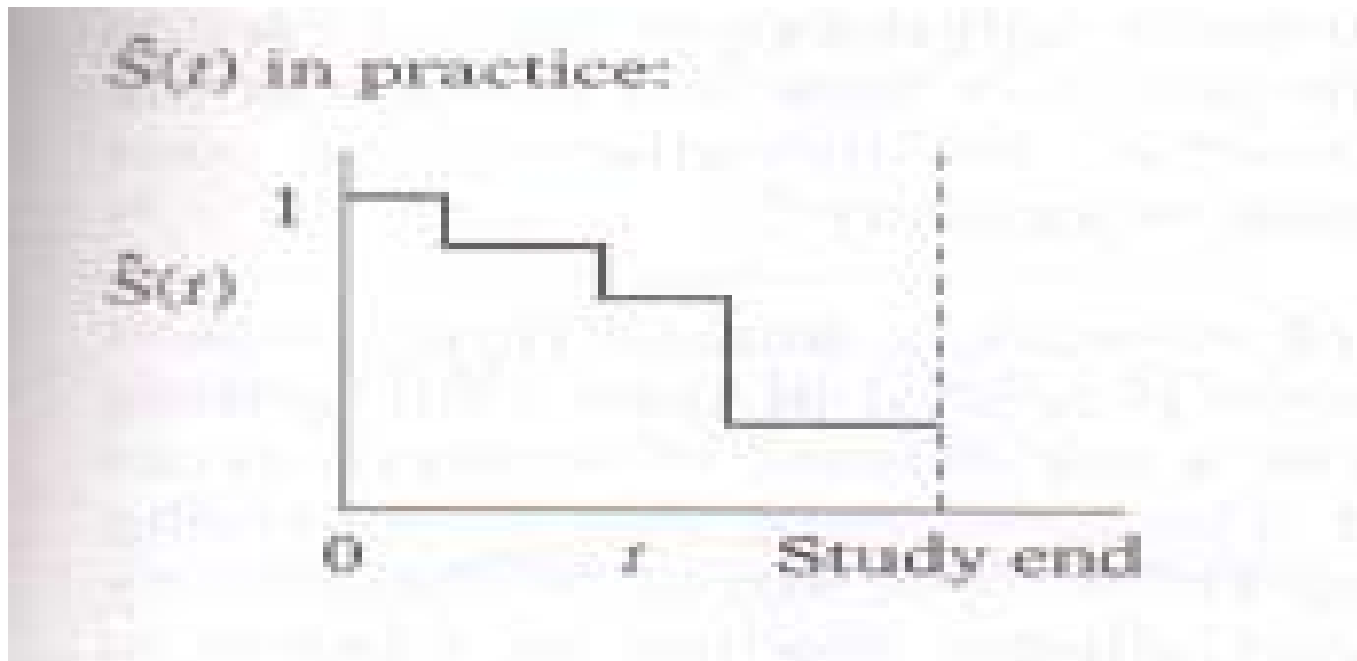
Survivor function

$S(t)$ = The probability that a subject survives longer than time t .

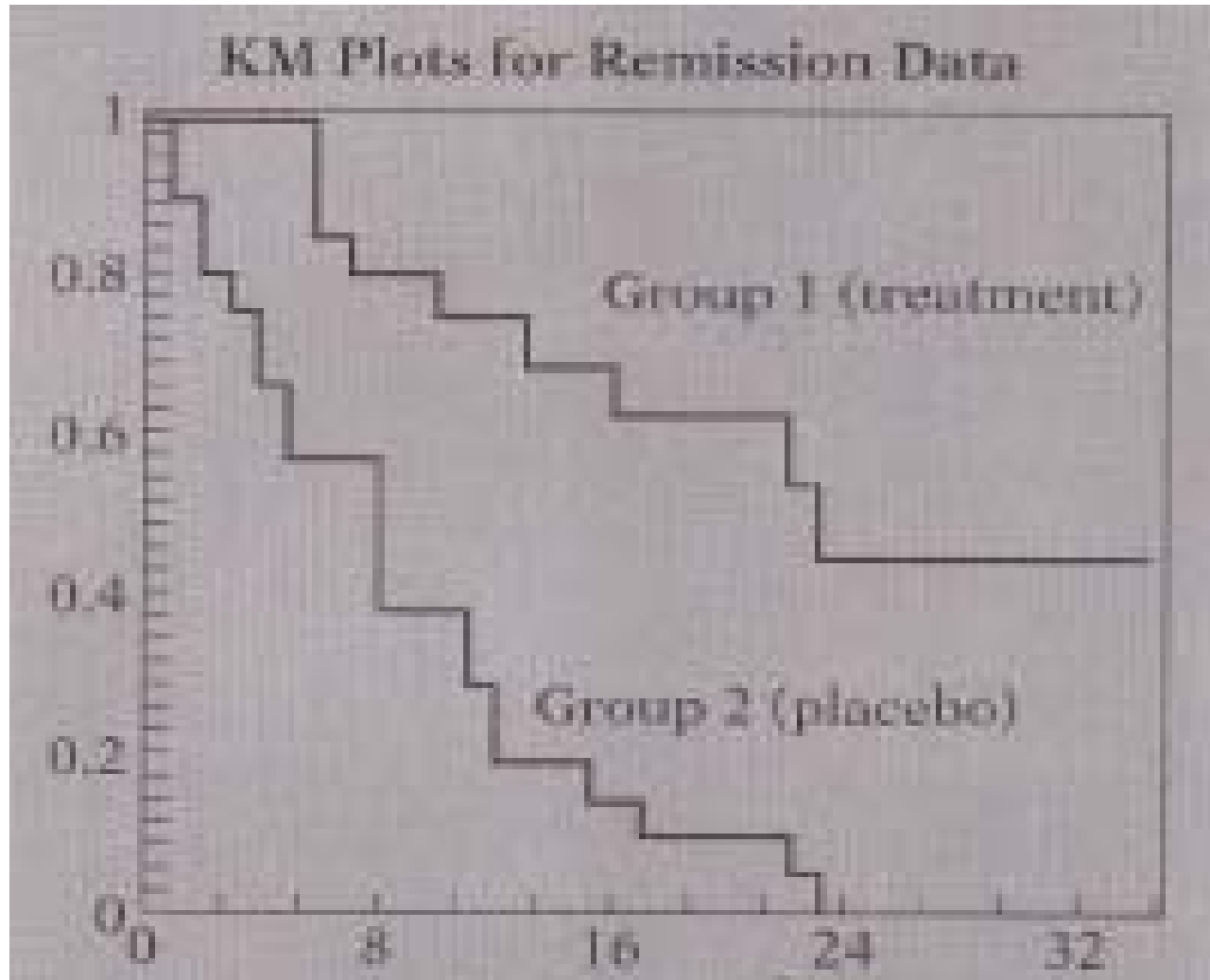


The survivor function is often expressed as a Kaplan-Meier curve

- Vertical drop indicates an event



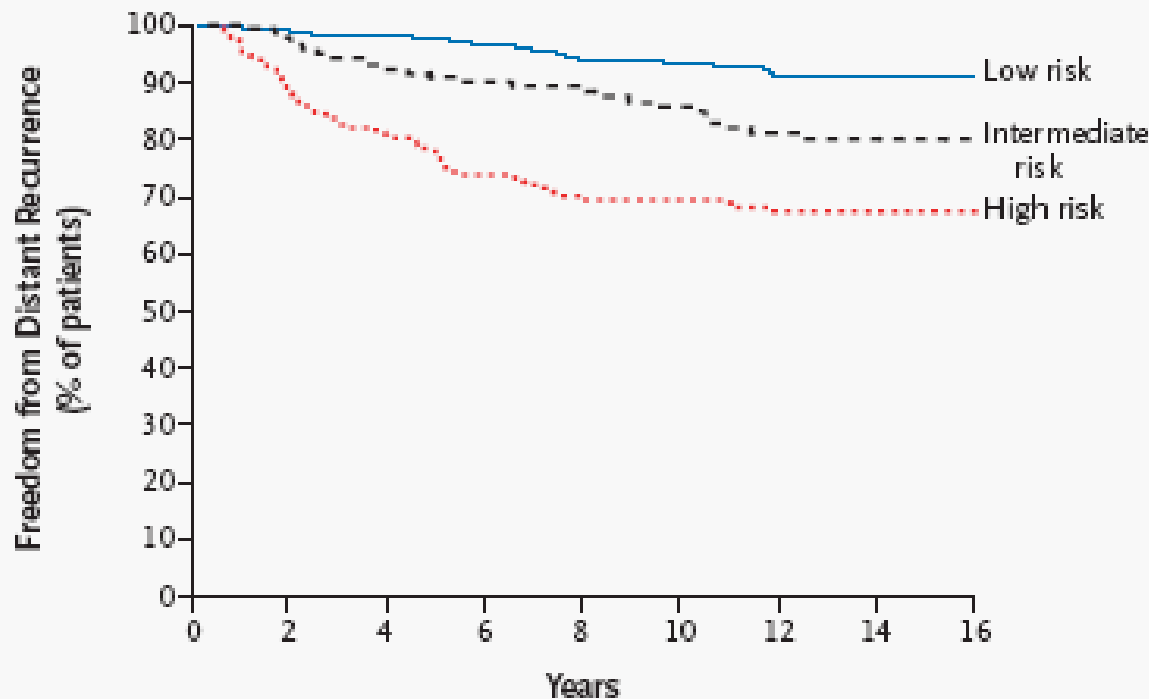
Kaplan-Meier survival curves



ORIGINAL ARTICLE

A Multigene Assay to Predict Recurrence of Tamoxifen-Treated, Node-Negative Breast Cancer

Soonmyung Paik, M.D., Steven Shak, M.D., Gong Tang, Ph.D.,
Chungyeul Kim, M.D., Joffre Baker, Ph.D., Maureen Cronin, Ph.D.,
Frederick L. Baehner, M.D., Michael G. Walker, Ph.D., Drew Watson, Ph.D.,
Taesung Park, Ph.D., William Hiller, H.T., Edwin R. Fisher, M.D.,
D. Lawrence Wickerham, M.D., John Bryant, Ph.D.,
and Norman Wolmark, M.D.



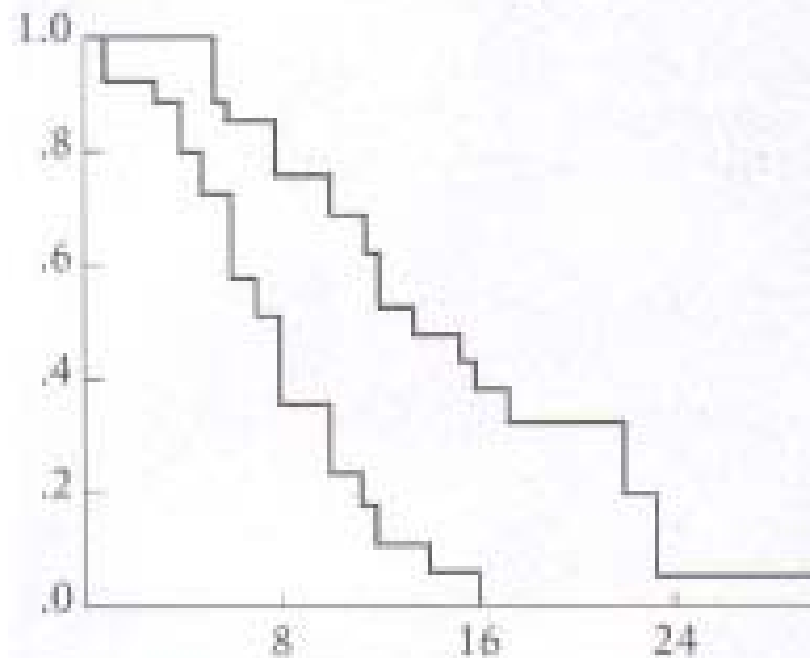
No. at Risk									
Low risk	338	328	313	298	276	258	231	170	38
Intermediate risk	149	139	128	116	104	96	80	66	16
High risk	181	154	137	119	105	91	83	63	13

Figure 2. Likelihood of Distant Recurrence, According to Recurrence-Score Categories.

A low risk was defined as a recurrence score of less than 18, an intermediate risk as a score of 18 or higher but less than 31, and a high risk as a score of 31 or higher. There were 28 recurrences in the low-risk group, 25 in the intermediate-risk group, and 56 in the high-risk group. The difference among the groups is significant ($P < 0.001$).

IV. The Log-Rank Test for Two Groups

Are KM curves statistically equivalent?



- Chi-square test
- Overall comparison of KM curves
- Observed versus expected counts
- Categories defined by ordered failure times

- Kaplan-Meier curves and the logrank test are useful when the predictor variable is categorical (e.g., drug vs. placebo) , or takes a small number of values (e.g., drug doses 0, 20, 50, and 100 mg/day) that can be considered to be categorical.
- The logrank test and K-M curves don't work easily with continuous predictors such as gene expression values.

- To do survival analysis using continuous variables such as gene expression or White Blood Count (WBC), we use Cox Proportional Hazards Regression Analysis (Cox PH regression).

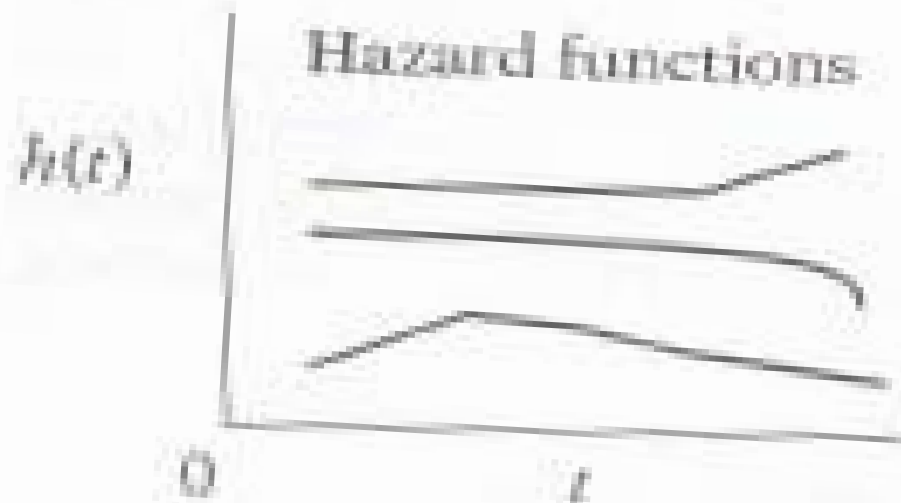
Hazard function $h(t)$

(1) A conceptual, approximate definition:

$h(t)$ is a function of the probability of an event in the time interval $[t, t+i]$, *given that the individual has survived up to time t*

(2) The formal calculus definition:

$$h(t) = \lim_{\Delta t \rightarrow 0} \frac{P(t \leq T < t + \Delta t \mid T \geq t)}{\Delta t}$$



- $h(t) \geq 0$
- $h(t)$ has no upper bound

Proportional hazards

- Consider two groups, A and B
- If the hazard function $h(t)$ for group A = $2 \cdot h(t)$ for group B, then the hazard rates for the two groups are proportional. The proportionality constant is 2. The hazards are proportional at all times t .

In general, if the hazard rate for one group is a constant multiple of the hazard for a second group, then the hazards are proportional.

Cox Proportional Hazards regression

- For continuous predictors such as gene expression values we can use Cox Proportional Hazard (PH) regression models.
- Cox PH can handle categorical data (e.g., treatment groups) by encoding as dummy $\{0, 1\}$ variables.

- Cox PH is a special type of survival analysis
 - handles censoring
- and a special type of regression analysis
 - handles continuous and categorical predictor variables

Cox proportional hazard regression

- To describe the effect on survival times of a continuous variable such as gene expression
 - Cox PH gives relative risk for a unit change in the variable
- E.g., a unit change in the expression of gene A gives a 2-fold increase in relative risk

Table 3. Multivariate Cox Proportional Analysis of Age, Tumor Size, Tumor Grade, and Recurrence Score in Relation to the Likelihood of Distant Recurrence.*

Variable	P Value	Hazard Ratio (95% CI)†
Analysis with recurrence score‡		
Age at surgery	0.22	0.76 (0.50–1.18)
Clinical tumor size	0.38	1.19 (0.81–1.76)
Tumor grade		
Moderately differentiated	0.15	1.55 (0.85–2.81)
Poorly differentiated	<0.001	3.34 (1.79–6.26)
HER2 amplification	0.06	0.51 (0.26–1.02)
Estrogen-receptor protein		
50–99 fmol/mg	0.32	0.75 (0.43–1.31)
100–199 fmol/mg	0.72	0.90 (0.52–1.58)
≥200 fmol/mg	0.94	1.02 (0.58–1.70)
Recurrence score	<0.001	2.81 (1.70–4.64)

I. A Computer Example Using the Cox PH Model

EXAMPLE

Leukemia Remission Data

Group 1($n = 21$)		Group 2($n = 21$)	
$t(\text{weeks})$	log WBC	$t(\text{weeks})$	log WBC
6	2.31	1	2.80
6	4.06	1	5.00
6	3.28	2	4.91
7	4.43	2	4.48
10	2.96	3	4.01
13	2.88	4	4.36
16	3.60	4	2.42
22	2.32	5	3.49
23	2.57	5	3.97
6+	3.20	8	3.52
9+	2.80	8	3.05
10+	2.70	8	2.32
11+	2.60	8	3.26
17+	2.16	11	3.49
19+	2.05	11	2.12
20+	2.01	12	1.50
25+	1.78	12	3.06
32+	2.20	15	2.30
32+	2.53	17	2.95
34+	1.47	22	2.73
35+	1.45	23	1.97

+ denotes censored observation

Model 1:

Column name	Coeff	StErr	p-value	HR	0.95	CI	P(PH)
Rx	1.509	0.410	0	4.523	2.027	10.094	0.794
n:42	%Cen: 28.571		-2 log L: 172.759				

Model 2:

Column name	Coeff	StErr	p-value	HR	0.95	CI	P(PH)
Rx	1.294	0.422	0.002	3.648	1.505	8.343	0.944
log WBC	1.604	0.329	0.000	4.975	2.609	9.486	0.917
n:42	%Cen: 28.571		-2 log L: 144.559				

- How does the Cox model work?

Hazard ratio (HR)

- Similar concept to odds ratio and relative risk
- $HR =$ The ratio of two hazard functions
- To get the Hazard for Group 2, *multiply* the Hazard for Group 1 by the Hazard ratio
- $HR = 4.5$ for Rx, meaning that the risk (of relapse) for group 2 is 4.5 times that of group 1.

- If $HR = 1$ then Group 1 $h(t) =$ Group 2 $h(t)$
- P-value for the Hazard Ratio is the probability that the $HR = 1$,
- P-value = 1 means that there is no difference in the hazard rates.
- P-value < 0.05 means a significant difference in the hazard rates.

II. The Formula for the Cox PH Model

$$h(t, \mathbf{X}) = h_0(t) e^{\sum \beta_j X_j}$$

$$\mathbf{X} = (X_1, X_2, \dots, X_p)$$

explanatory/predictor variables

II. The Formula for the Cox PH Model

$$h(t, \mathbf{X}) = h_0(t) e^{\sum \beta_i X_i}$$

$$\mathbf{X} = (X_1, X_2, \dots, X_p)$$

explanatory/predictor variables

The Cox PH model is usually written in terms of the hazard model formula shown here at the left. This model gives an expression for the hazard at time t for an individual with a given specification of a set of explanatory variables denoted by the bold \mathbf{X} . That is, the bold \mathbf{X} represents a collection (sometimes called a "vector") of predictor variables that is being modeled to predict an individual's hazard.

$h_0(t)$

\times

$$\sum_{i=1}^p \beta_i X_i$$

Baseline hazard

Involves t but
not X 's

Exponential

Involves X 's but
not t (X 's are time-
independent)

- In the Cox model, when all the X 's are zero, the formula reduces to the baseline hazard, $h_0(t)$

$$\begin{aligned} h(t, \mathbf{X}) &= h_0(t) e^{\sum \beta_j X_j} \\ &= h_0(t) e^0 \\ &= h_0(t) \end{aligned}$$

Baseline hazard

No X 's in model: $h(t, \mathbf{X}) = h_0(t)$.

- In a Cox PH model, we can determine the significant coefficients (variables) without specifying the baseline hazard.
- Use maximum likelihood (ML) to determine coefficients
- Given a Cox model and the coefficients, we can subsequently estimate the baseline hazard function and the survival curves.

IV. ML Estimation of the Cox PH Model

$$h(t, \mathbf{X}) = h_0(t) e^{\sum_{i=1}^p \beta_i x_i}$$

ML estimates: $\hat{\beta}_i$

Column name	Coeff	StErr	p-value	HR
Rx	1.294	0.422	0.002	3.648
log WBC	1.604	0.329	0.000	4.975
n=42	%Cen: 28.571		-2 log L: 144.559	

Model 2:

$$h(t, \mathbf{X}) = h_0(t) e^{\beta_1 Rx + \beta_2 \log WBC}$$

Estimated model:

$$\hat{h}(t, \mathbf{X}) = \hat{h}_0(t) e^{1.294 Rx + 1.604 \log WBC}$$

- Cox Proportional Hazards regression analysis assumes that the hazards are proportional (constant ratio) over time.
- If the hazards are not proportional, then the model is wrong, and conclusions are likely to be wrong.
- Software does checks for proportional hazards assumption.

- Cox models can also be used with time-varying covariates, such as gene expression values that change over time.