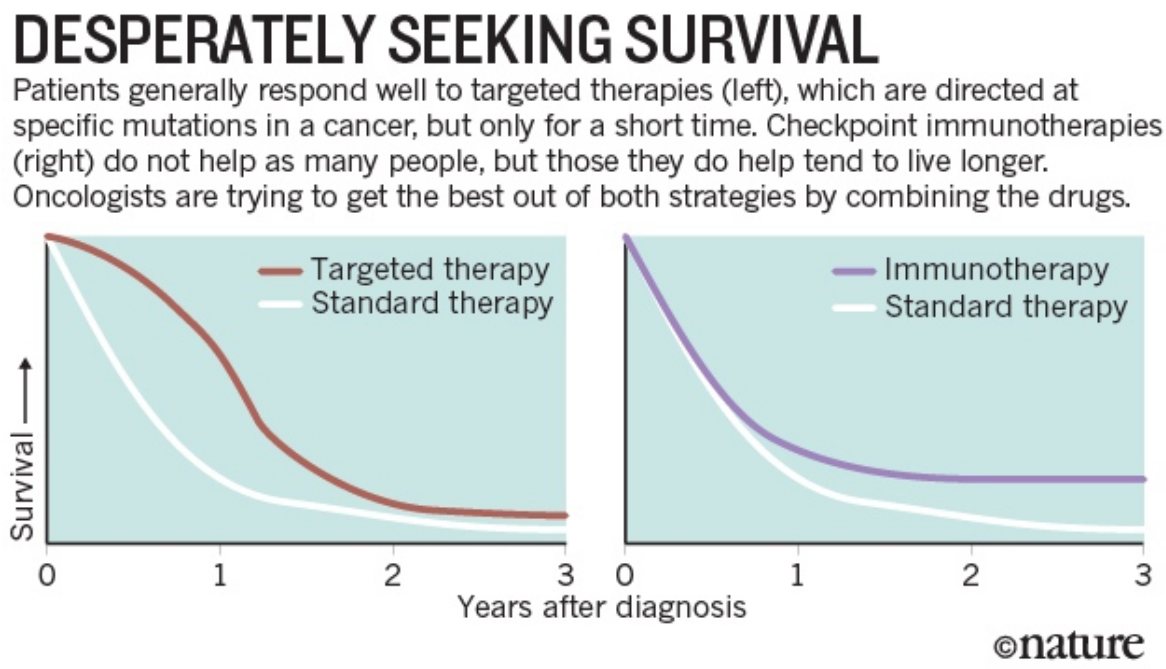
Survival Analysis

DSLA COURSE

ROHIT PADEBETTU

Survival Analysis



Everything exists in time

a.k.a.

Time-to-Event Analysis

Clinical Trials – Drug Efficacy

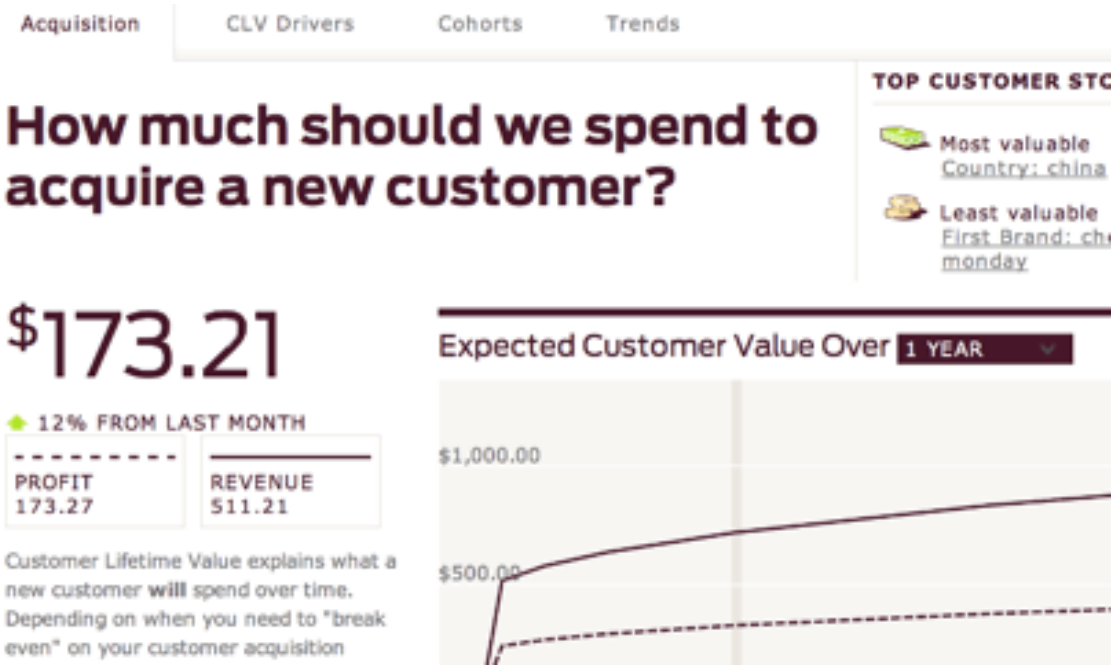
Abiraterone acetate

ble 3: Overall Survival of Patients Treated with Either ZYTIGA or Placebo in Combination with Prednisone (Intent-to-Treat Analysis)

	ZYTIGA (N=797)	Placebo (N=398)
Primary Survival Analysis		
Deaths (%)	333 (42%)	219 (55%)
Median survival (months) (95% CI)	14.8 (14.1, 15.4)	10.9 (10.2, 12.0)
p value [a]	< 0.0001	
Hazard ratio (95% CI) [b]	0.646 (0.543, 0.768)	
Updated Survival Analysis		
Deaths (%)	501 (63%)	274 (69%)
Median survival (months) (95% CI)	15.8 (14.8, 17.0)	11.2 (10.4, 13.1)
Hazard ratio (95% CI) [b]	0.740 (0.638, 0.859)	

"Survival Analysis" is a technique long used in health sciences

Customer Life Time Value



It enables asking questions like

How much time
before an event occurs,
if it occurs?

Survival Analysis

A *set of statistical techniques* where the outcome variable is the *time until occurrence of an event* of interest

Applications

Reliability Analysis in Engineering
Failure in Mechanical Systems

Default Risk Analysis
Understand which customers are more likely to default and by when

Customer Churn Analysis in Education & Business
Understand which customers are likely to switch, drop off and when

Time to lapsing of policy
in Insurance

Event History Analysis
in Sociology
like occurrence of disease, death, marriage, divorce

Time events to critical events
in Public Sector

Churn vs Survival

Churn Analysis

- **Examines customer churn within a set time window e.g. next 3 or 6 months**
- **Predicts likelihood of customer to churn during the defined window**
- **No indication about subsequent risk of churn**
- **Does not provide information on customer lifetime value**

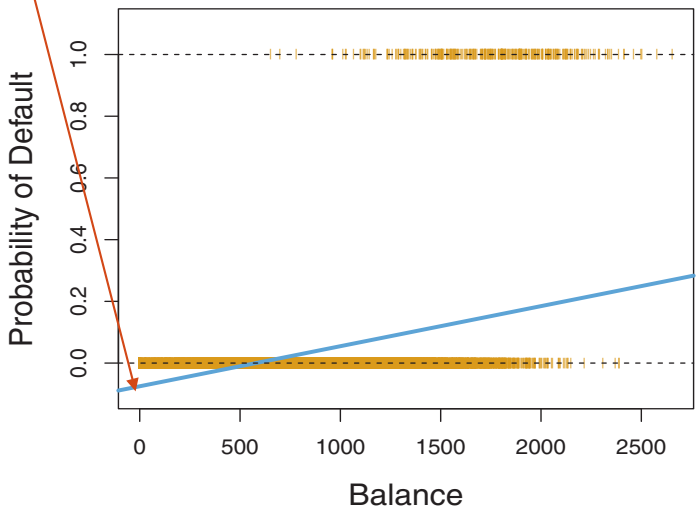
Survival Analysis

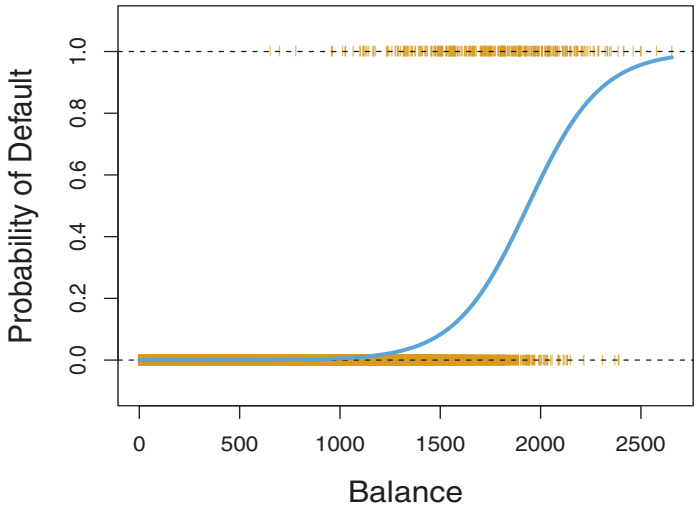
- **Examines how churn takes place over time**
- **Describes or predicts retention likelihood over time**
- **Identifies key points in customer lifecycle**
- **Informs customer lifetime value**

Linear Regression -> Logistic Regression

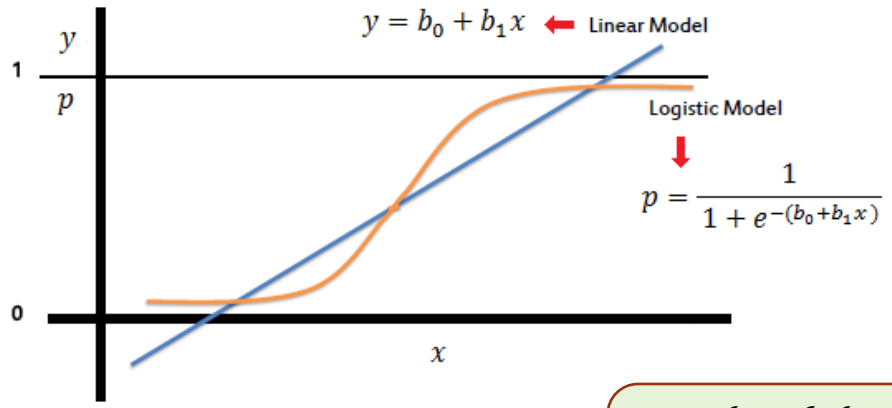
Can the Probability be less than Zero ?

Loan Default Probability





Logistic Regression



$$\log\left(\frac{p(X)}{1-p(X)}\right) = \beta_0 + \beta_1 X.$$

Here we model the

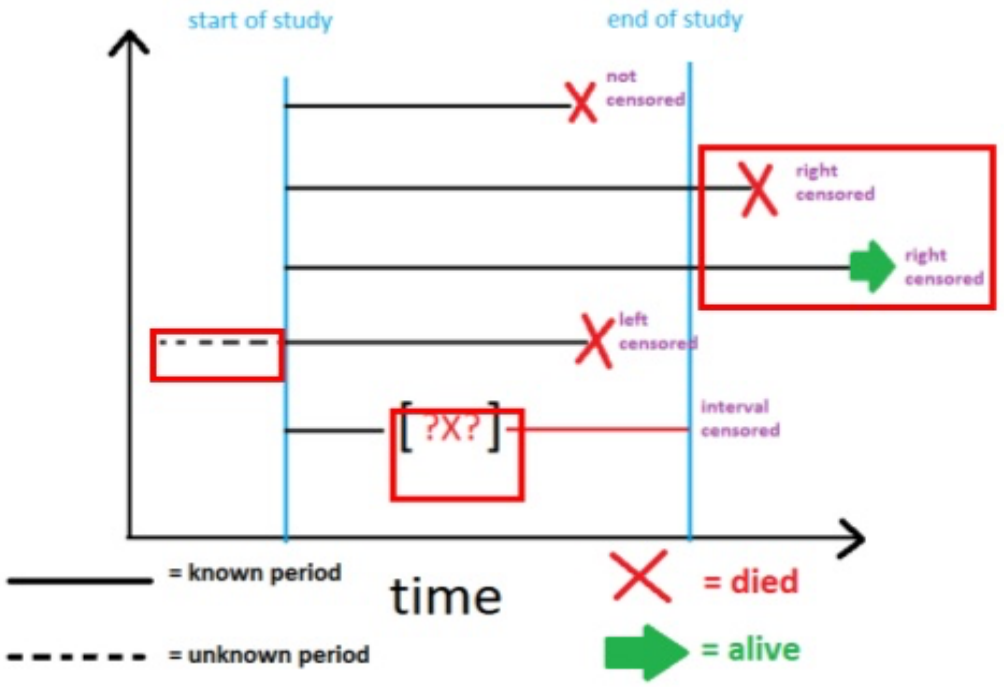
Success/Failure Ratio
Or
"Odds Ratio"

How do we model time effect?

What if the data were incomplete/censored as in Clinical Trials?

What is Censoring?

What is Censoring in Data Science?



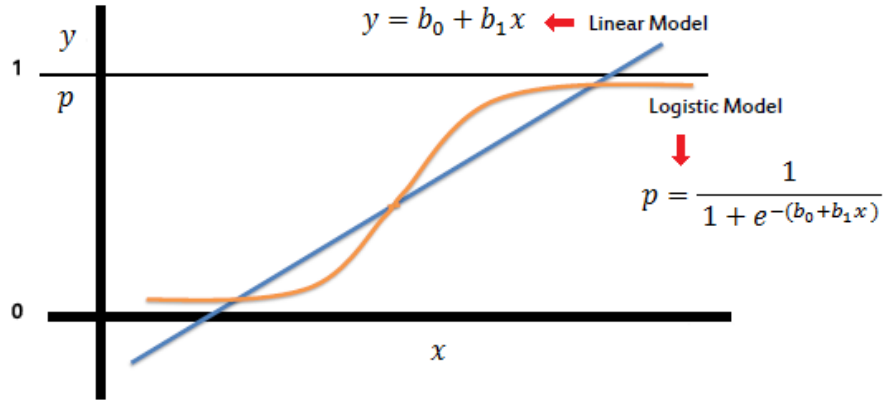
Right censoring is most common in defined period studies

What if customer leaves next day?

What if person has a heart attack next day?

What if the car never comes for its first service?

Logistic Regression

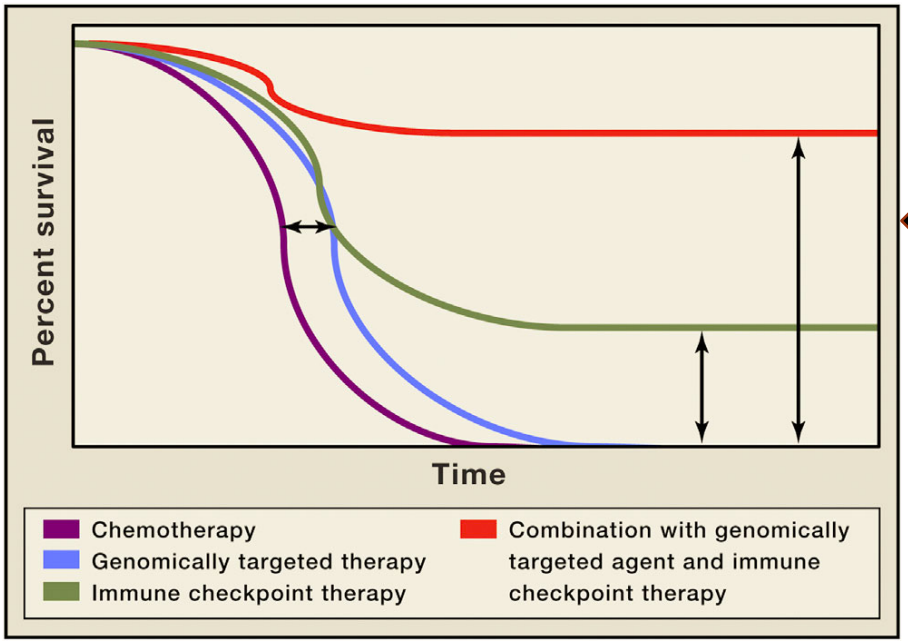


$$\log\left(\frac{p(X)}{1-p(X)}\right) = \beta_0 + \beta_1 X.$$

Logistic Regression would give a biased prediction if Censoring isn't accounted for

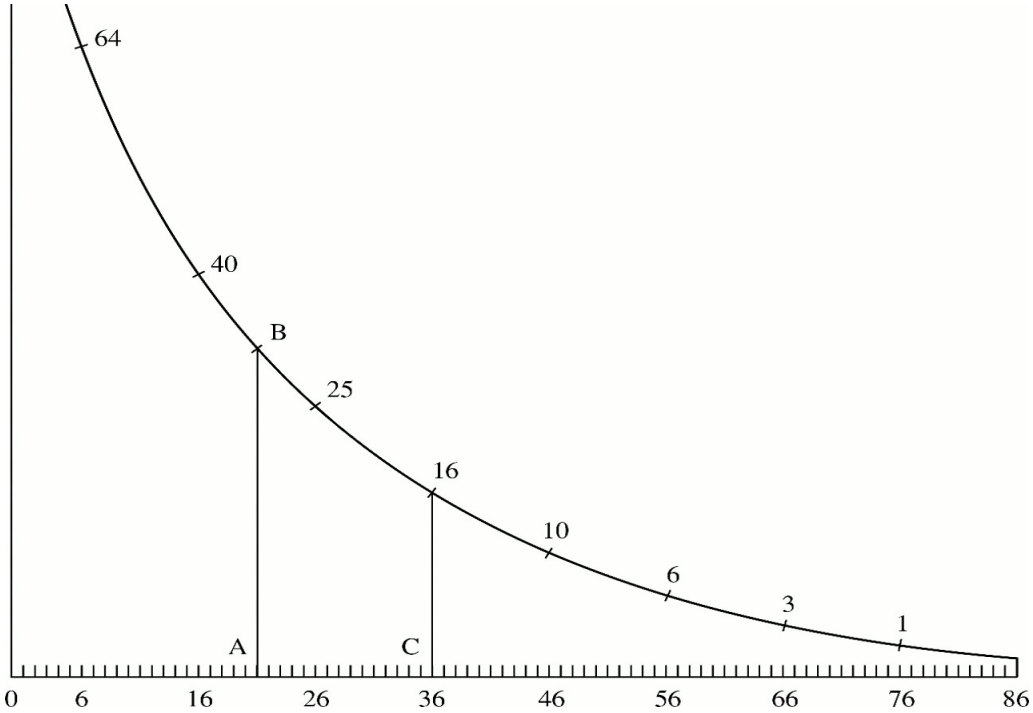
Survival Analysis

Survival Curve



Survival Curve of patients undergoing different forms of Cancer Therapy

What is Survival Analysis?

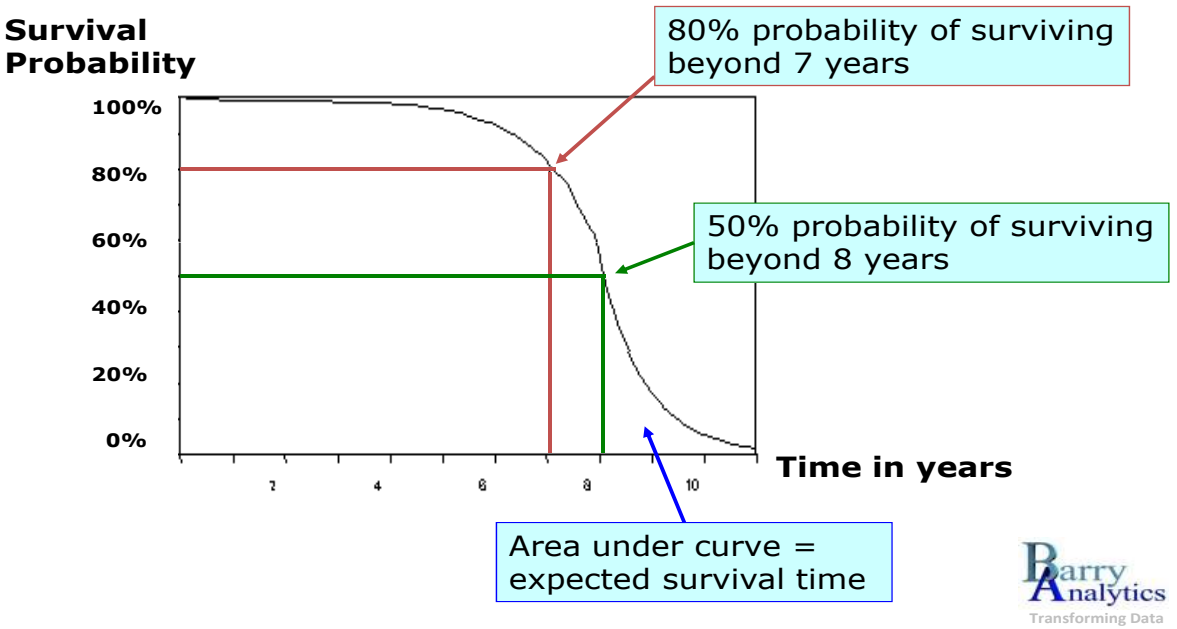


Roughly, what shape is this function?

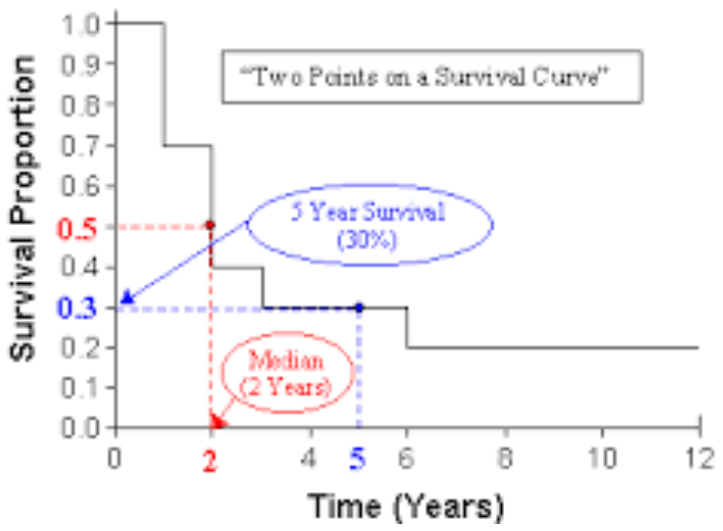
What was a person's chance of surviving past 20? Past 36?

This is survival analysis! We are trying to estimate this curve—only the outcome can be any binary event, not just death.

Descriptive Survival Analysis

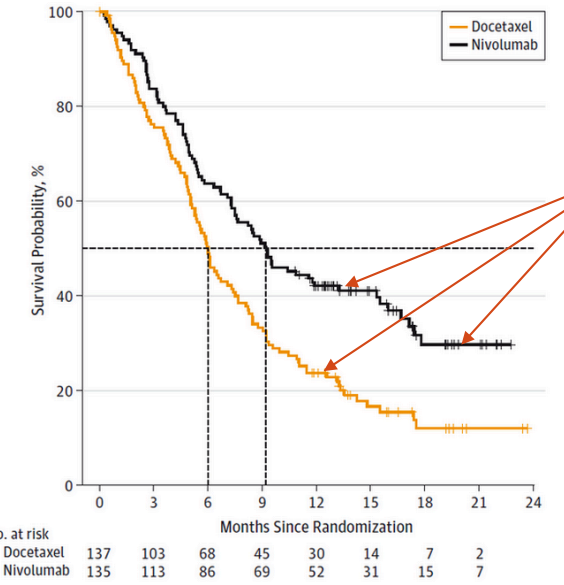


Survival Curves



$$S(t) = \frac{number\ of\ individuals\ surviving\ longer\ than\ t}{total\ number\ of\ individuals\ studied}$$

Figure. Kaplan-Meier Curves of Overall Survival Found in Study CM017

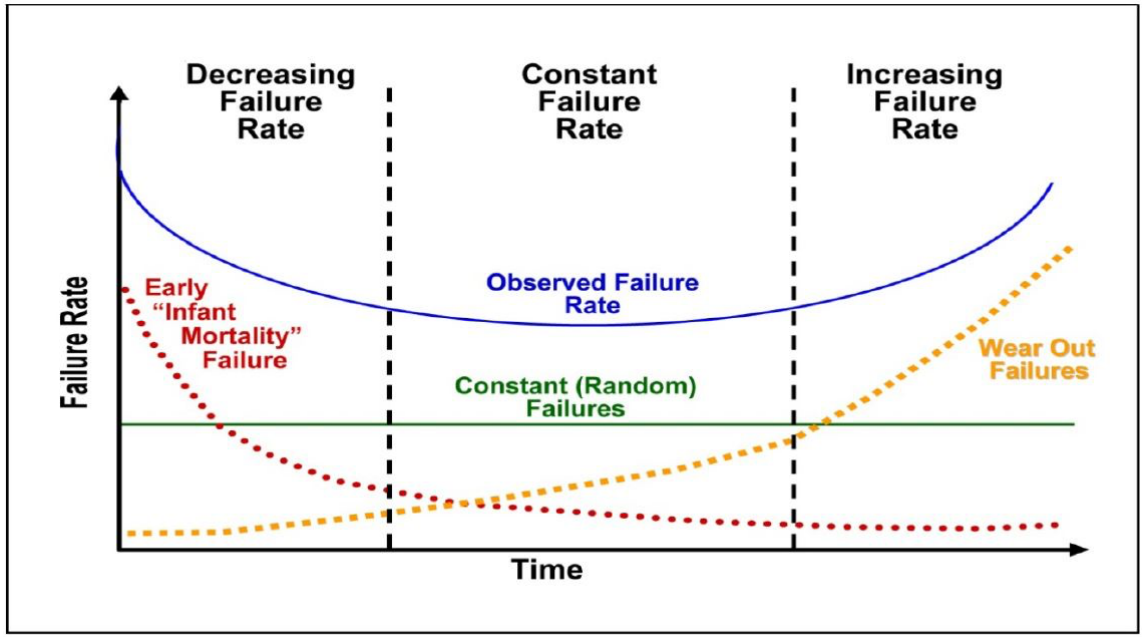


The CM017 study is well described by Brahmer and colleagues.[18] Per the O'Brien-Fleming boundary,[19] the significance level for the interim overall survival analysis with 199 deaths was 2-sided $P = .03$.

Censoring

Failure Rate

Failure = 1-Survival
Failure Rate = Slope of that Failure Curve



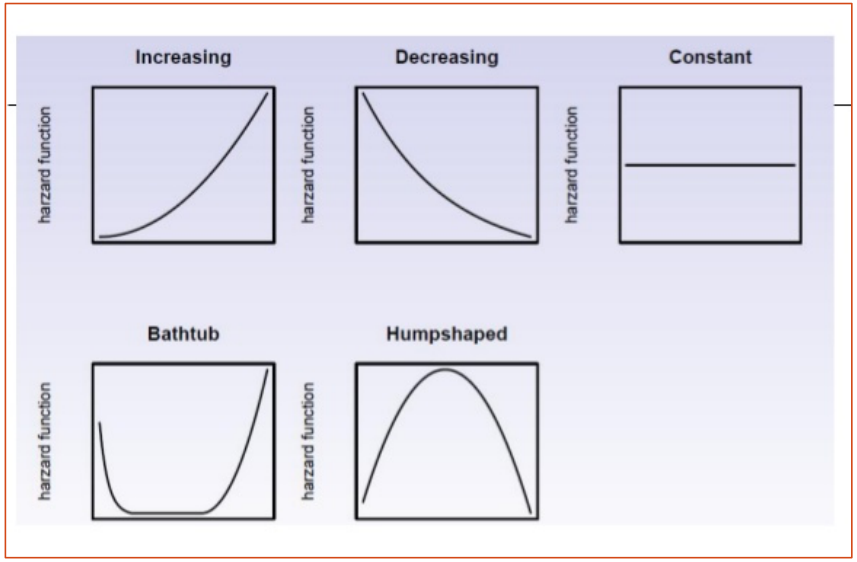
Hazard Rates

*The probability that **if you survive to time 't'**, you will succumb to the event in the next instant.*

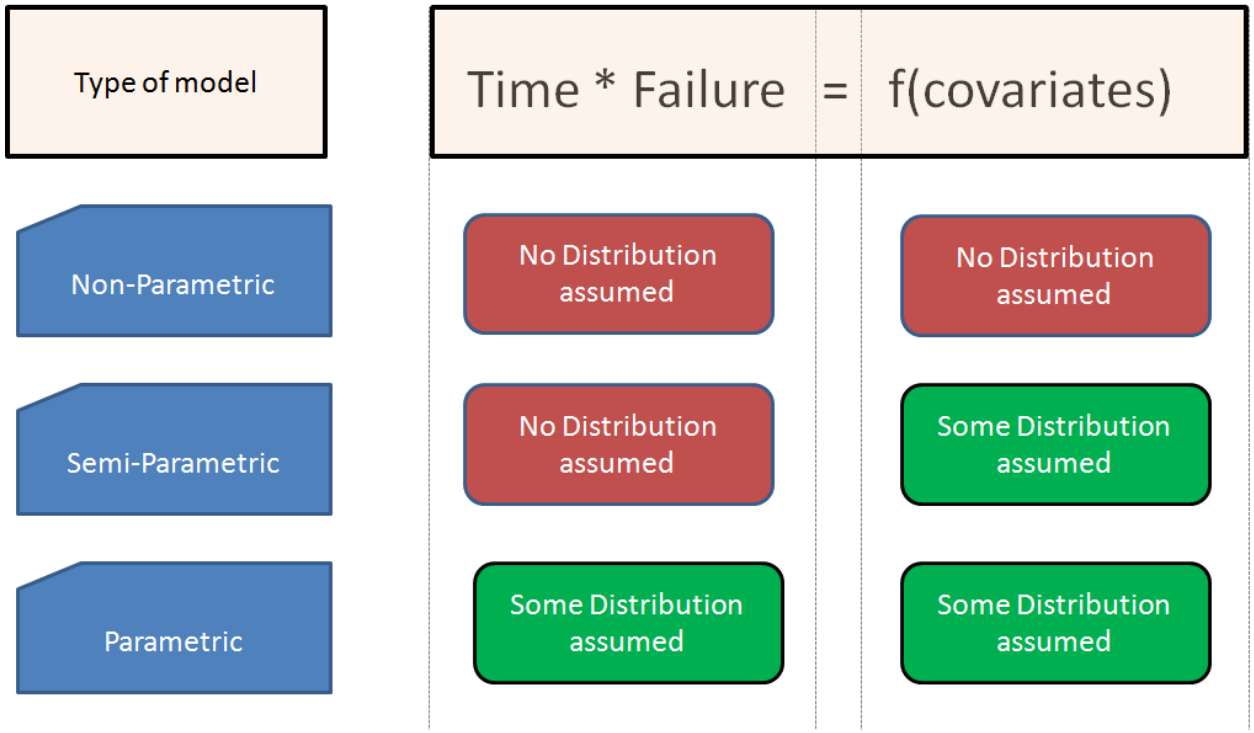
It is the instantaneous rate of the event happening

Slope of the Survival Curve

$$h(t) = -\frac{S'(t)}{S(t)}$$



Hazard Models

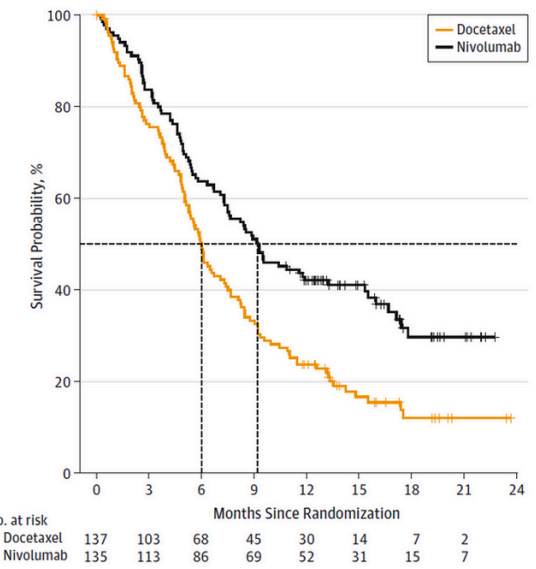


Kaplan Meier – Non Parametric Model

- *Non-parametric estimate of the survival function.*
- *Commonly used to describe survivorship of study population*
- *Commonly used to compare two study populations.*
- *Intuitive graphical presentation.*

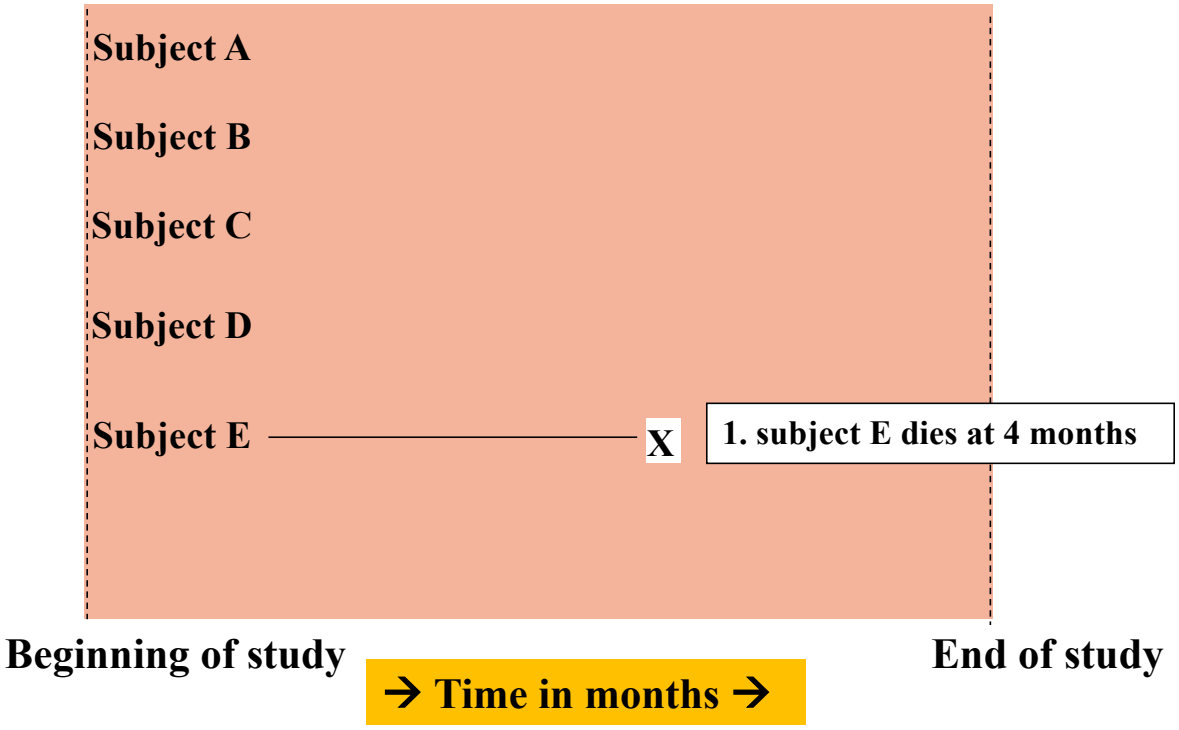
$$S(t) = \frac{number\ of\ individuals\ surviving\ longer\ than\ t}{total\ number\ of\ individuals\ studied}$$

Figure. Kaplan-Meier Curves of Overall Survival Found in Study CMO17

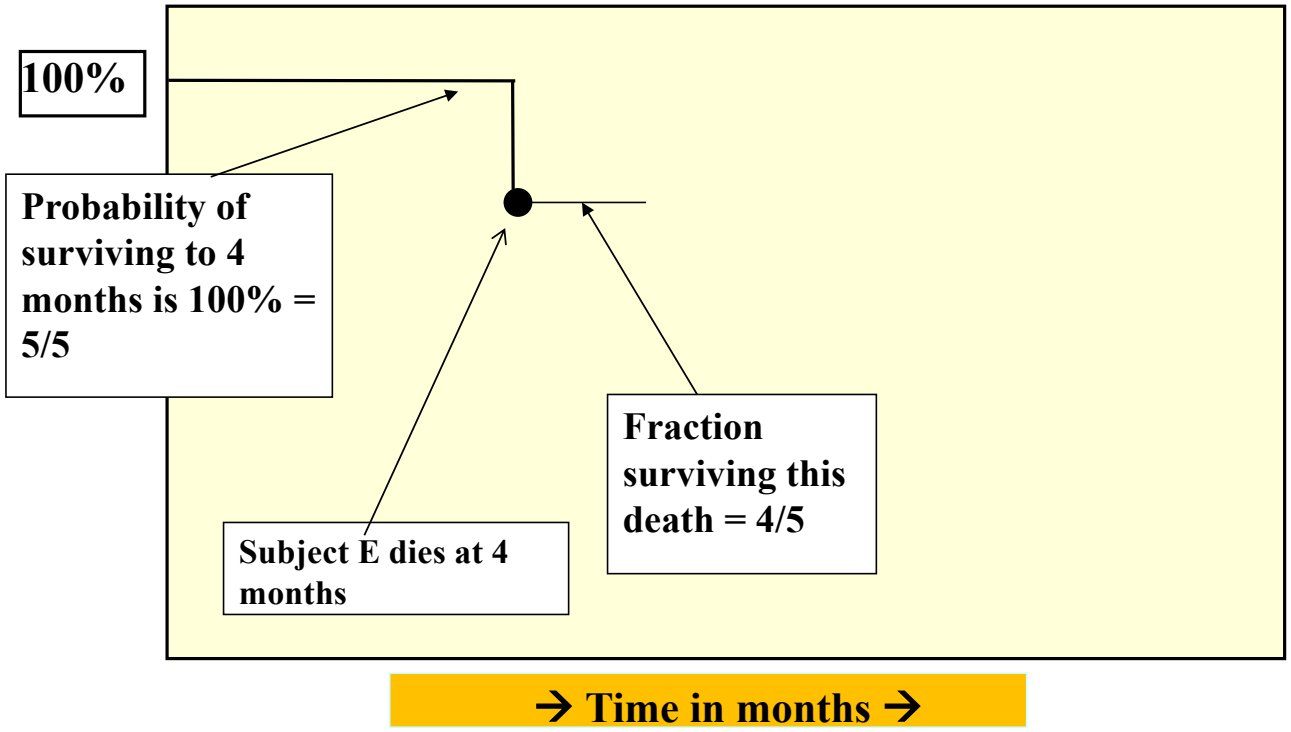


The CMO17 study is well described by Brahmer and colleagues.[18] Per the O'Brien-Fleming boundary,[19] the significance level for the interim overall survival analysis with 199 deaths was 2-sided $P = .03$.

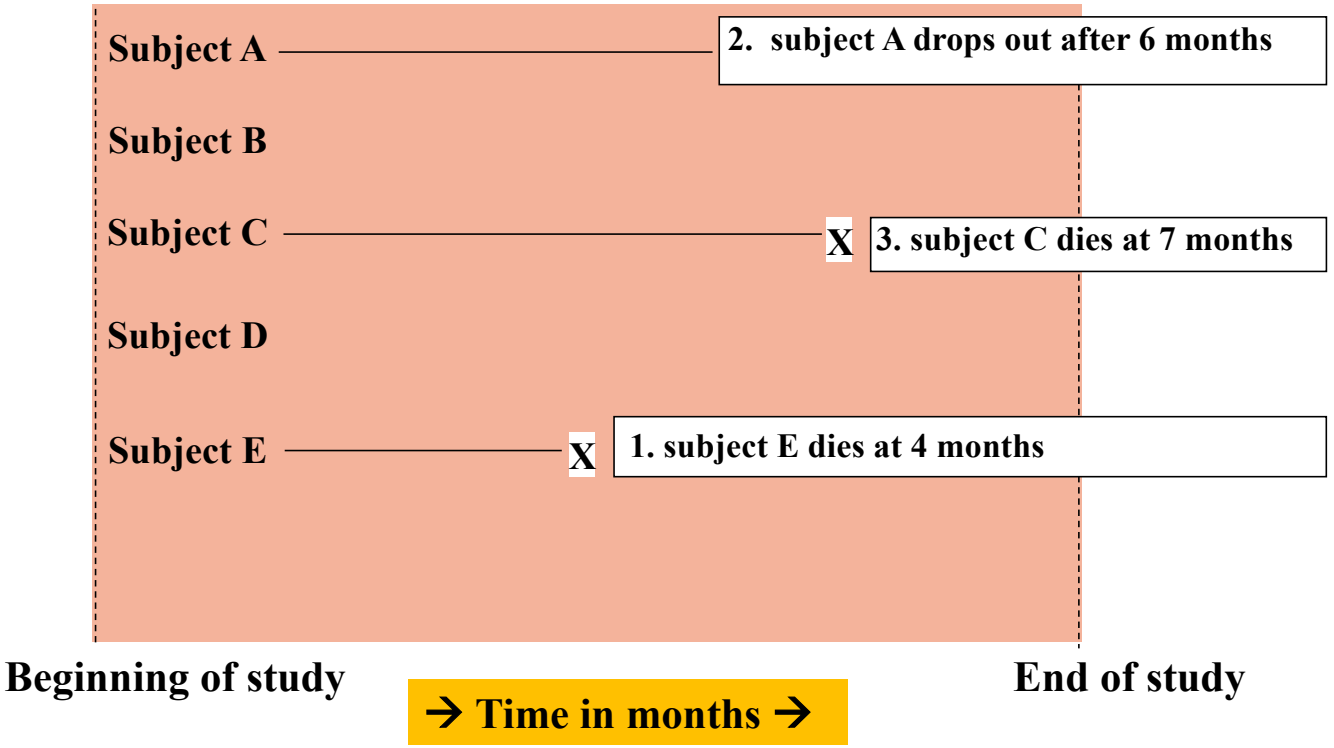
Survival Data (right-censored)



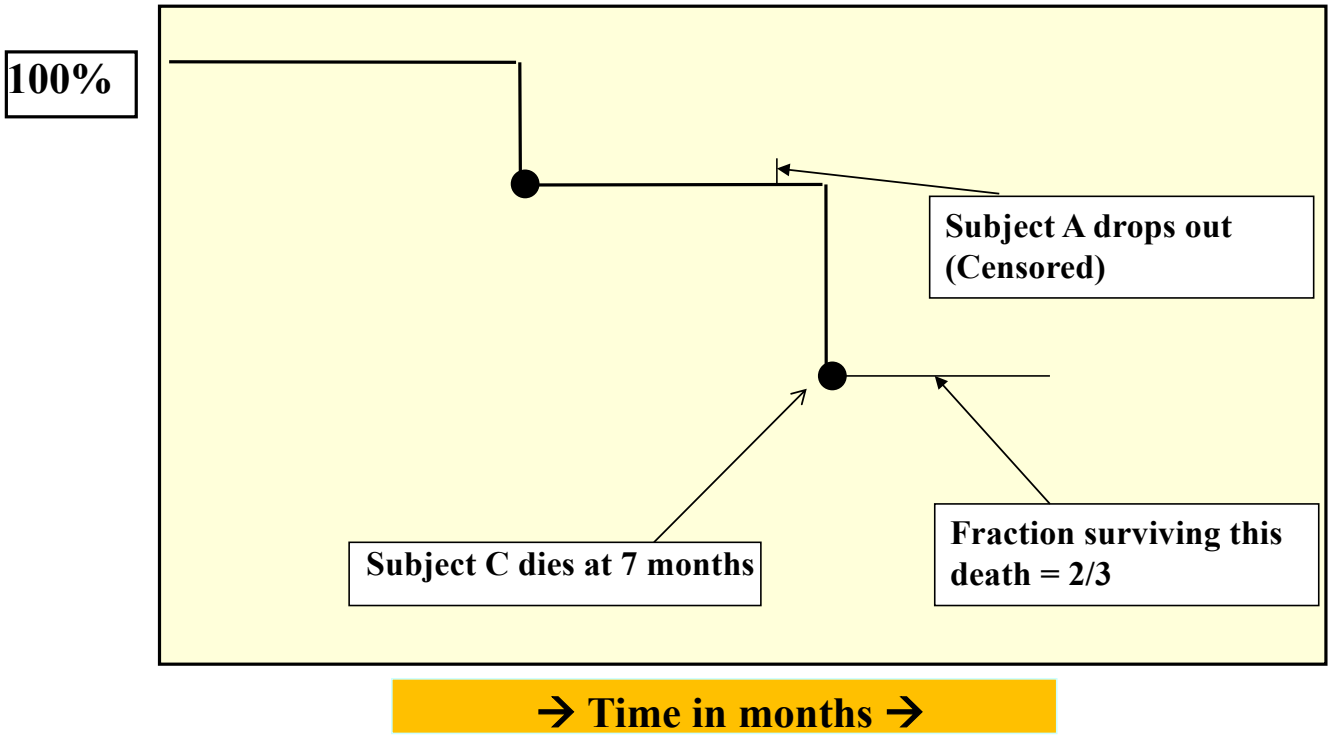
Corresponding Kaplan-Meier Curve



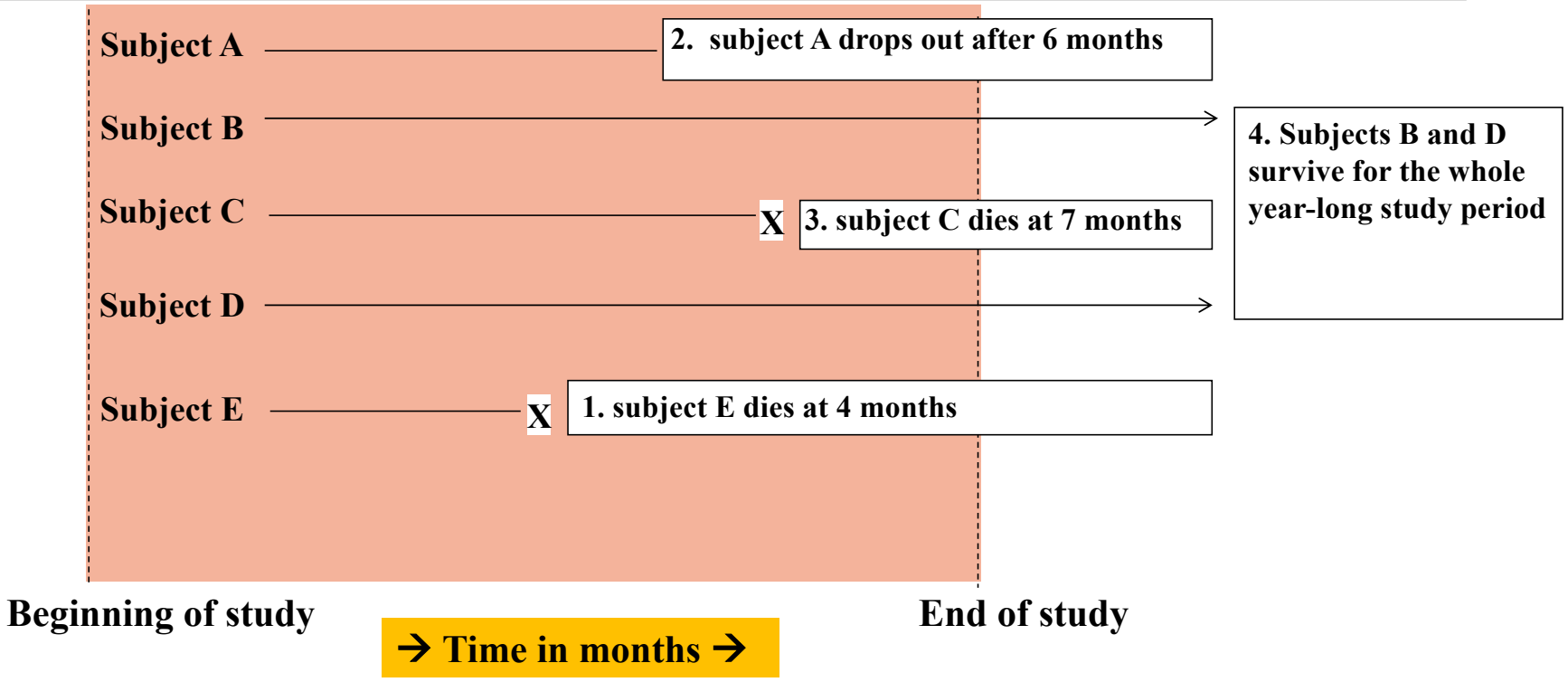
Survival Data



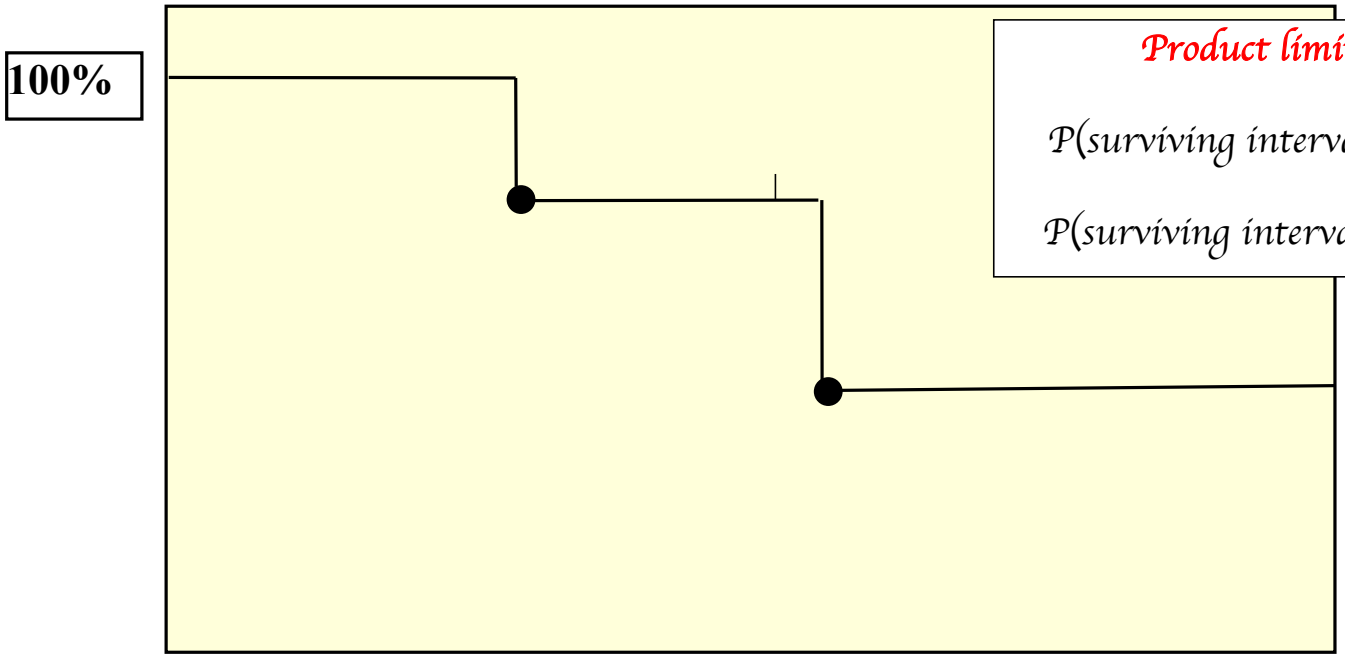
Corresponding Kaplan-Meier Curve



Survival Data



Corresponding Kaplan-Meier Curve



Product limit estimate of survival

P(surviving interval 1/at-risk up to failure 1)
*
P(surviving interval 2/at-risk up to failure 2)

*4/5 * 2/3 = 0.5333*

→ Time in months →

Kaplan-Meier Model

Limitations

- Mainly descriptive
- Doesn't control for covariates
- Requires categorical predictors
- Can't accommodate time-dependent variables

Benefits

- No math assumptions about the hazard model.
- Suitable for univariate analysis
- Simply, the empirical probability of surviving past certain times in the sample (taking into account censoring).

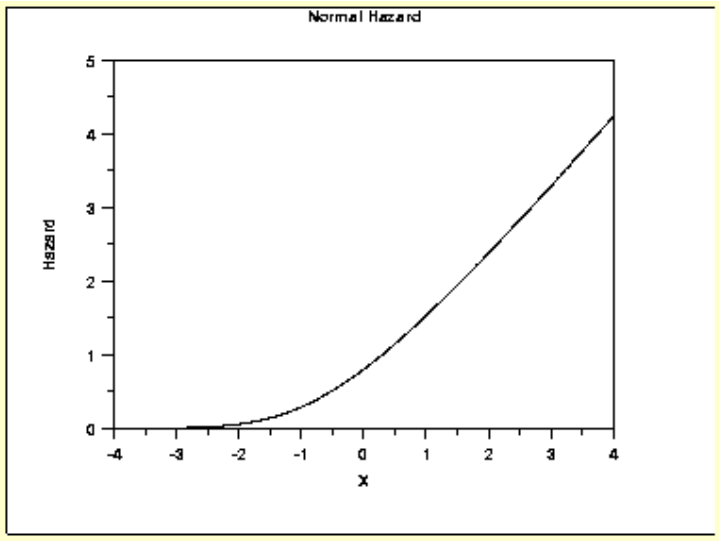
Parametric Hazard Models

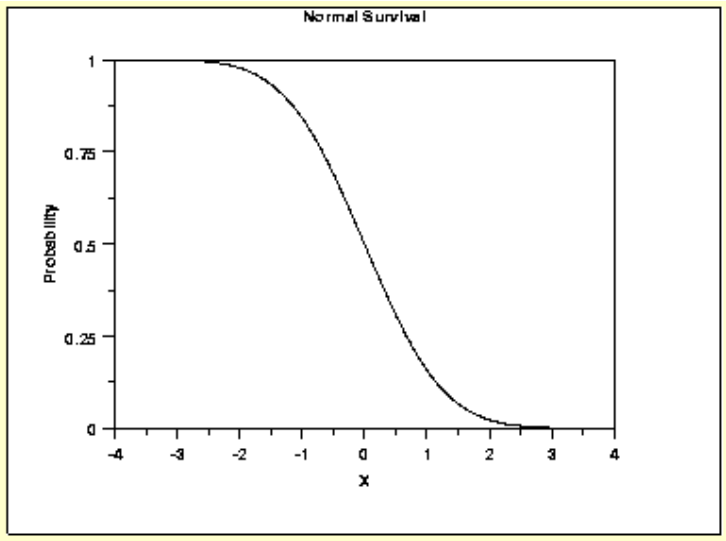
- Model the underlying hazard/survival function
- Assume that the dependent variable (time-to-event) takes on some known distribution, such as Weibull, exponential, or lognormal.
- Estimates parameters of these distributions (e.g., baseline hazard function)
- Estimates covariate-adjusted hazard ratios.
  - A hazard ratio is a ratio of hazard rates

2 Components

- A baseline hazard function (which may change over time)
- A linear function of a set of k fixed covariates that when exponentiated gives the relative risk.

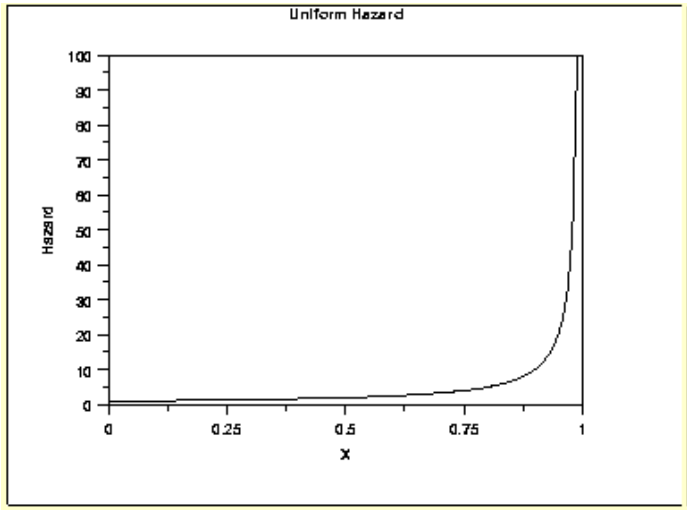
Normal Hazard Model

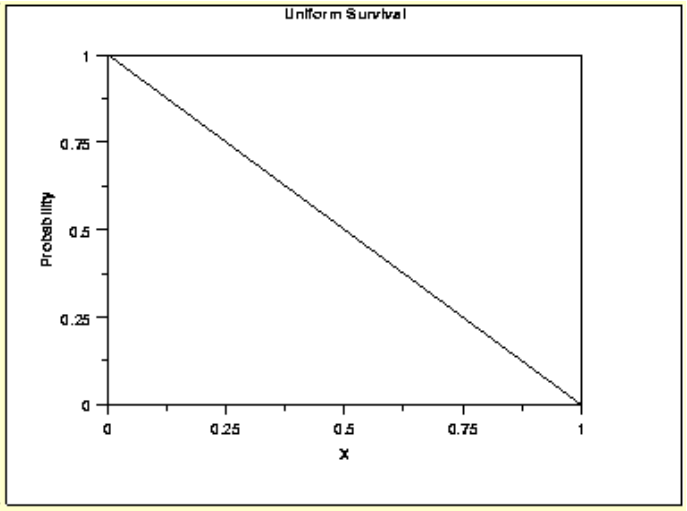




Assumes the risk of failure increases considerably with time
Ex: Mechanical systems like car

Uniform Hazard Model



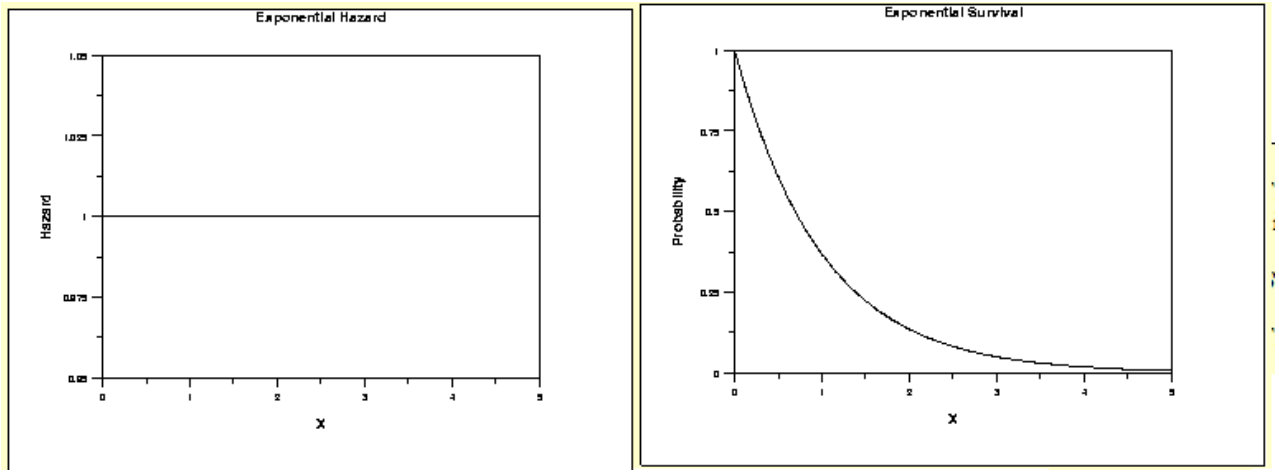


Risk of failure increases exponentially leading to constant decline in Survival. Not common in real world

Exponential Hazard Model

Exponential model assumes fixed baseline hazard that we can estimate.

$$\log h_i(t) = \mu + \beta_1 x_{i1} + \ldots + \beta_k x_{ik}$$

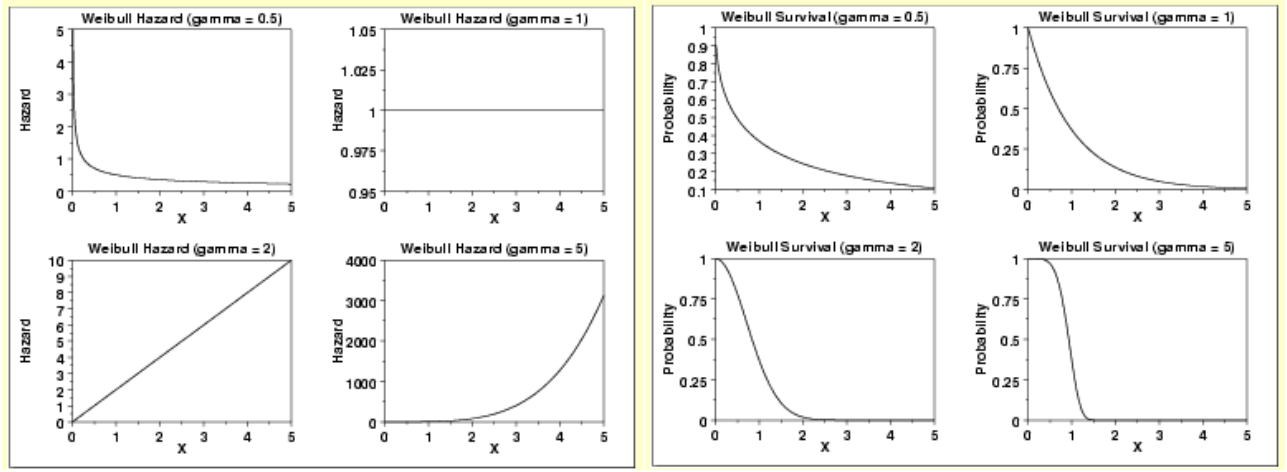


A Constant Hazard Rate is assumed, leading to an exponential decline in survival ex: Lifespan of humans

Weibull Hazard Model

Weibull model models the baseline hazard as a function of time. Two parameters (shape and scale) must be estimated to describe the underlying hazard function over time.

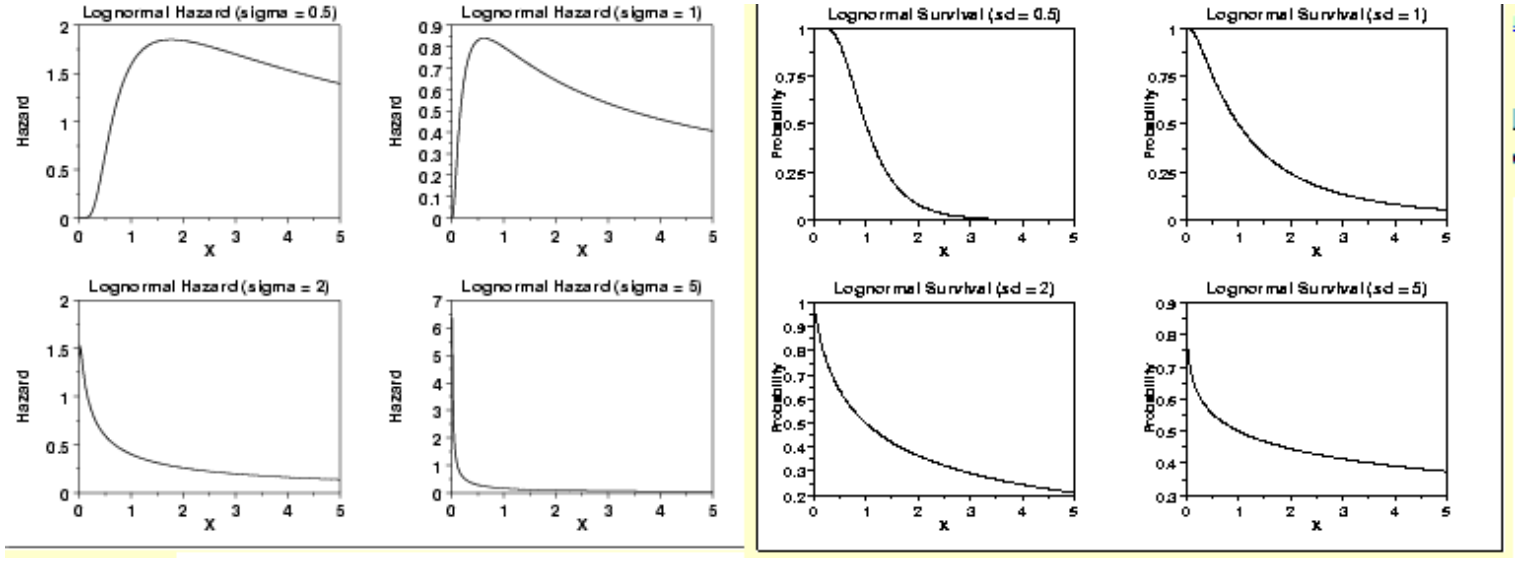
$$\log h_i(t) = \mu + \alpha \log t + \beta_1 x_{i1} + \ldots + \beta_k x_{ik}$$



A gamma rate can be adjusted to model a monotonically trending Hazard Rate. Ex: Cancer Survival as risk increases with time

Log-normal Hazard Model

A sigma rate can be adjusted to model a non-monotonically trending Hazard Rate ex: Mortgage Loan Defaults



Semi Parametric Models

Many times we care more about comparing groups than about estimating absolute survival

A very popular and widely used model

Cox Regression

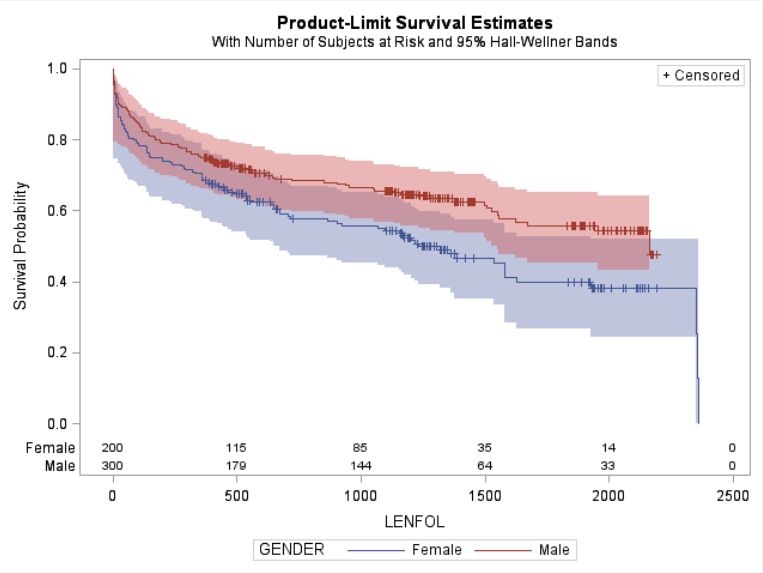
"A Proportional Hazards Model"

Allows Multivariate Analysis

$$\log h_i(t) = \log h_0(t) + \beta_1 x_{i1} + \ldots + \beta_k x_{ik}$$

$$h_i(t) = h_0(t) e^{\beta_1 x_{i1} + \ldots + \beta_k x_{ik}}$$

No need to decide this!



Cox Proportional Hazard

The point is to compare hazard rates of groups who have different values for predictors

To estimate **relative risk** rather than **absolute risk**

$$HR = \frac{h_1(t)}{h_2(t)} = \frac{h_0(t)e^{\beta x_1}}{h_0(t)e^{\beta x_2}} = e^{\beta(x_1 - x_2)}$$

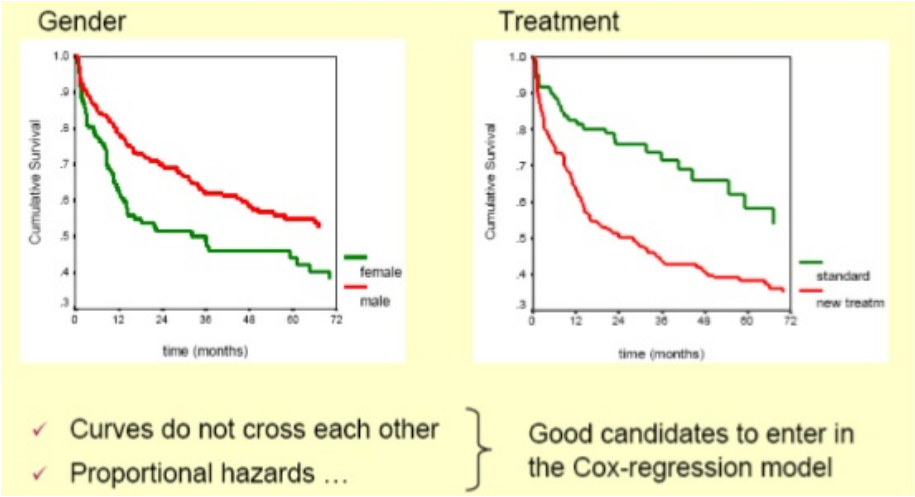
Independent of Time

HR = 1: No effect
HR < 1: Reduction in the hazard
HR > 1: Increase in Hazard

In cancer studies:

- A covariate with hazard ratio > 1 (i.e.: $b > 0$) is called *bad prognostic factor*
- A covariate with hazard ratio < 1 (i.e.: $b < 0$) is called *good prognostic factor*

Cox Proportional Hazard




```
> summary(coxph(srv.dt2~groups))
Call:
coxph(formula = srv.dt2 ~ groups)

  n= 250, number of events= 168

           coef exp(coef) se(coef)      z Pr(>|z|)
groupsB -0.2921    0.7467   0.1953 -1.495    0.135
groupsC -1.2477    0.2872   0.1991 -6.267 3.69e-10 ***

Likelihood ratio test= 44.03  on 2 df,   p=2.752e-10
Wald test            = 40.2  on 2 df,   p=1.861e-09
Score (logrank) test = 43.69  on 2 df,   p=3.261e-10
```

Hazard Ratios: *26% decline in risk for Group B compared to Group A*

Survival Analysis in R

Demo

Survival Analysis

Have good rest of weekend!