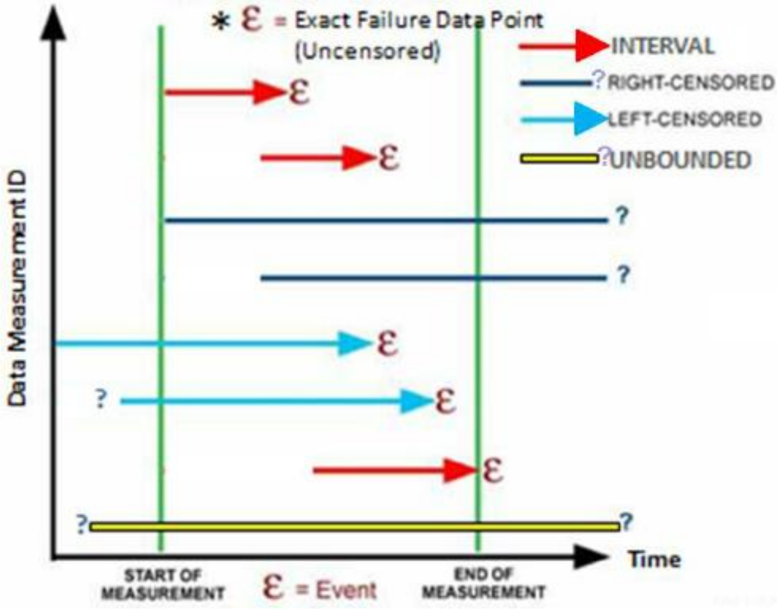
Data Analytics

Survival Analysis



Dr. Rita Chakravarti
Institute of Systems Science
National University of Singapore
Email: rita@nus.edu.sg

Topics

- Introduction – What is Survival Analysis?
- What is censoring?
- The Hazard Function
- Cox's Proportional Hazards Model
- The Kaplan-Meier Estimator

Introduction - What is Survival Analysis?

- Survival analysis is just another name for time to event analysis.
- The term survival analysis is used predominately in biomedical sciences where the interest is in observing time to death either of patients or of laboratory animals (an example of such a data set is provided in the next slide)
- Time to event analysis has also been used widely in
  - Medical studies, where issues like survival time, time to remission, etc.... are studied
  - The social sciences where interest is on analyzing time to events such as job changes, marriage, birth of children etc...
  - The engineering sciences have also contributed to the development of survival analysis which is called "reliability analysis" or "failure time analysis" (Dependability, or **reliability**, describes the ability of a system or component to function under stated conditions for a specified period of time)
  - Where the focus is on modeling the time it takes for machines or electronic components (or software components) to break down.
- In summarizing survival data, there are two functions of central interest, namely the *survivor function and the hazard function.*

Medical Science Data – WHAS100

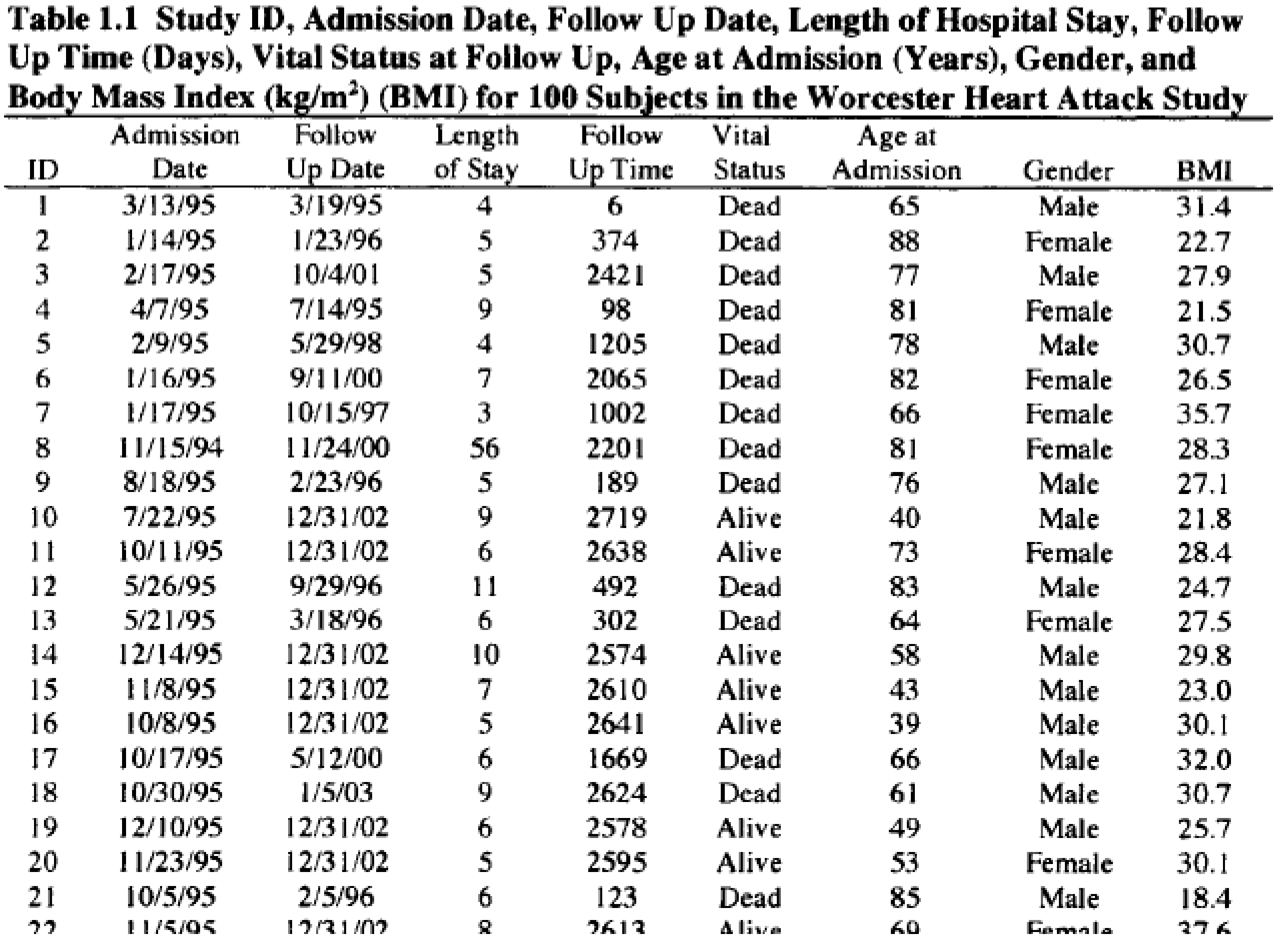


Table 1.1 Study ID, Admission Date, Follow Up Date, Length of Hospital Stay, Follow Up Time (Days), Vital Status at Follow Up, Age at Admission (Years), Gender, and Body Mass Index (kg/m^2) (BMI) for 100 Subjects in the Worcester Heart Attack Study

ID	Admission Date	Follow Up Date	Length of Stay	Follow Up Time	Vital Status	Age at Admission	Gender	BMI
1	3/13/95	3/19/95	4	6	Dead	65	Male	31.4
2	1/14/95	1/23/96	5	374	Dead	88	Female	22.7
3	2/17/95	10/4/01	5	2421	Dead	77	Male	27.9
4	4/7/95	7/14/95	9	98	Dead	81	Female	21.5
5	2/9/95	5/29/98	4	1205	Dead	78	Male	30.7
6	1/16/95	9/11/00	7	2065	Dead	82	Female	26.5
7	1/17/95	10/15/97	3	1002	Dead	66	Female	35.7
8	11/15/94	11/24/00	56	2201	Dead	81	Female	28.3
9	8/18/95	2/23/96	5	189	Dead	76	Male	27.1
10	7/22/95	12/31/02	9	2719	Alive	40	Male	21.8
11	10/11/95	12/31/02	6	2638	Alive	73	Female	28.4
12	5/26/95	9/29/96	11	492	Dead	83	Male	24.7
13	5/21/95	3/18/96	6	302	Dead	64	Female	27.5
14	12/14/95	12/31/02	10	2574	Alive	58	Male	29.8
15	11/8/95	12/31/02	7	2610	Alive	43	Male	23.0
16	10/8/95	12/31/02	5	2641	Alive	39	Male	30.1
17	10/17/95	5/12/00	6	1669	Dead	66	Male	32.0
18	10/30/95	1/5/03	9	2624	Dead	61	Male	30.7
19	12/10/95	12/31/02	6	2578	Alive	49	Male	25.7
20	11/23/95	12/31/02	5	2595	Alive	53	Female	30.1
21	10/5/95	2/5/96	6	123	Dead	85	Male	18.4
22	11/5/95	12/31/02	8	2613	Alive	60	Female	37.6

Introduction - What is the Survival Function?

- When we are considering the length of time a person will live, or the length of time a machine will operate before breakdown then our primary interest is the **survival function**, denoted by *S(t)*, which is defined by

  $$S(t) = \text{Probability } (T > t)$$

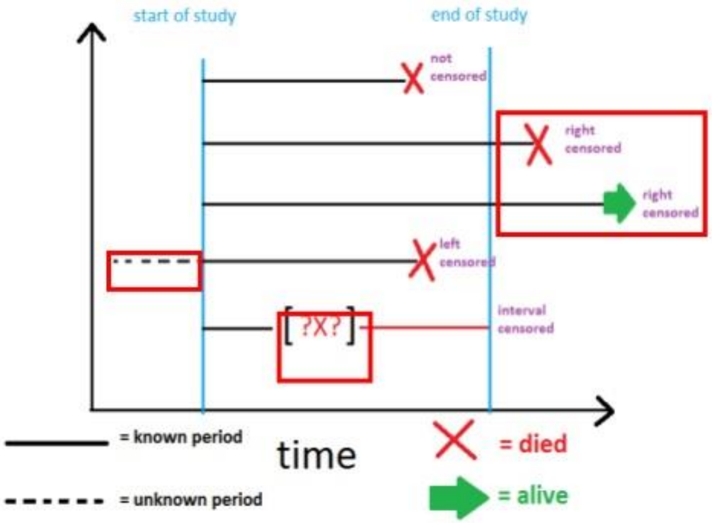
  - where *t* is some time measured after the person is born, or the machine starts operating ,
  - *T* is a random variable denoting the time of death/ breakdown (an event occurs)
- That is, the survival function is the probability that the time of death is later than some specified time *t*.
  - Usually one assumes $S(0) = 1$, although it could be less than 1 if there is the possibility of immediate death or failure.
  - The survival function is usually assumed to approach zero as age increases without bound, i.e., $S(t) \to 0$ as $t \to \infty$

Introduction - What is Survival Analysis?

- The **lifetime distribution function**, conventionally denoted *F*, is defined as the complement of the survival function $F(t) = \Pr(T \leq t) = 1 - S(t).$
  - It is usually a **non-normal** distribution
- There are certain aspects of survival analysis data, such as censoring and non-normality, that generate great difficulty when trying to analyze the data using traditional statistical models such as multiple linear regression.
- The non-normality aspect of the data violates the normality assumption of most commonly used statistical models and tests such as regression or ANOVA, etc.
- A censored observation is defined as an observation with incomplete information.
  - Example: you may track an individual from birth to a certain age, but then loose track of him

Introduction: Censoring

- Censoring is a form of missing data problem which is common in survival analysis.
- Ideally, both the birth and death dates of a subject are known,
  - In which case the lifetime is known.
- Suppose we are carrying out a study over a period of time.
- There are four different types of censoring possible:
  - Right truncation,
  - Left truncation,
  - Right censoring
  - Left censoring



What is Censoring and Truncation?

- If a subject's lifetime is known to be less than a certain duration, the lifetime is said to be *left-censored*.
- It may also happen that subjects with a lifetime less than some threshold may not be observed at all: this is called *truncation*.
- Note that truncation is different from left censoring, since
  - for a left censored datum, we know the subject exists,
  - but for a truncated datum, we may be completely unaware of the subject.

  Truncation is also common.
- In a so-called *delayed entry* study, subjects are not observed at all until they have reached a certain age.
- For example, people may not be observed until they have reached the age to enter school. Any deceased subjects in the pre-school age group would be unknown.
- Left-truncated data are common in actuarial work for life insurance and pensions.

What is Censoring? (cont.)

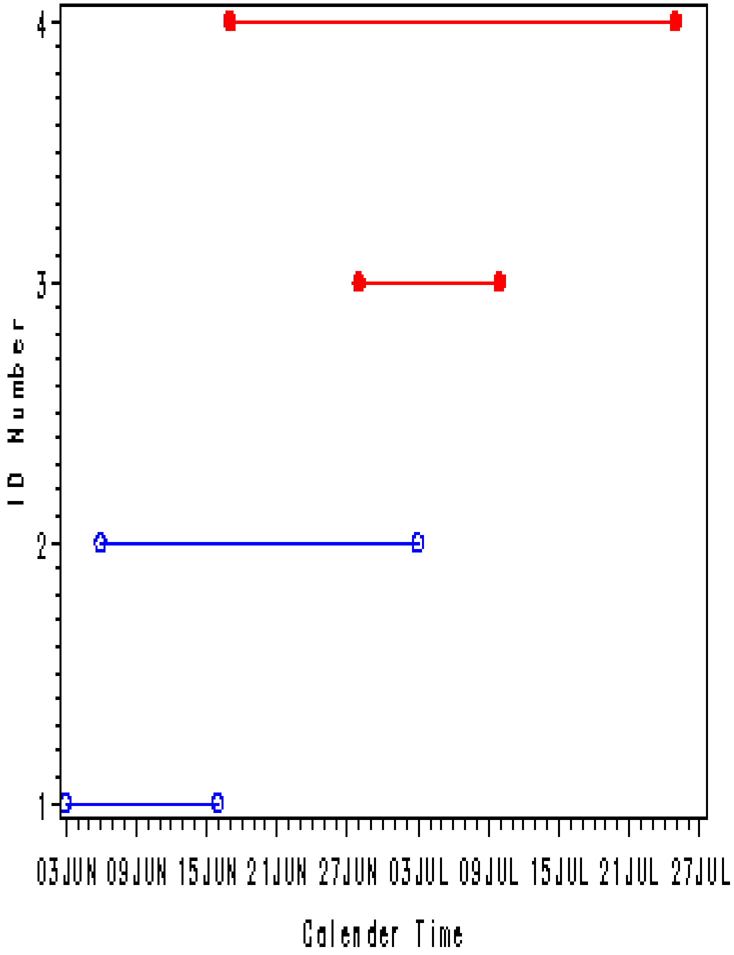
- We generally encounter right-censored data.
- Left-censored data can occur when a person's survival time becomes incomplete on the left side of the follow-up period for the person.
- As an example, we may follow up a patient for any infectious disorder from the time of his or her being tested positive for the infection.
  - ⇨ We may never know the exact time of exposure to the infectious agent.
- We will focus exclusively on right censoring for a number of reasons.
  - – Most data used in analyses have only right censoring.
  - – Furthermore, right censoring is the most easily understood of all the four types of censoring and if a researcher can understand the concept of right censoring thoroughly it becomes much easier to understand the other three types.
  - – When an observation is right censored it means that the information is incomplete because the subject did not have an event during the time that the subject was part of the study.

Impact of Censoring

- The aim of survival analysis is to follow subjects over time
  - And observe at which point in time they experience the event of interest (probably death, or recurrence of a medical condition or failure of a machine)
- It often happens that the study does not span enough time in order to observe the event for all the subjects in the study.
- This could be due to a number of reasons. Perhaps subjects drop out of the study for reasons unrelated to the study
  - E.g. patients moving to another area and leaving no forwarding address).
- The common feature of all of these examples is that if the subject had been able to stay in the study then it would have been possible eventually to observe the time of the event.

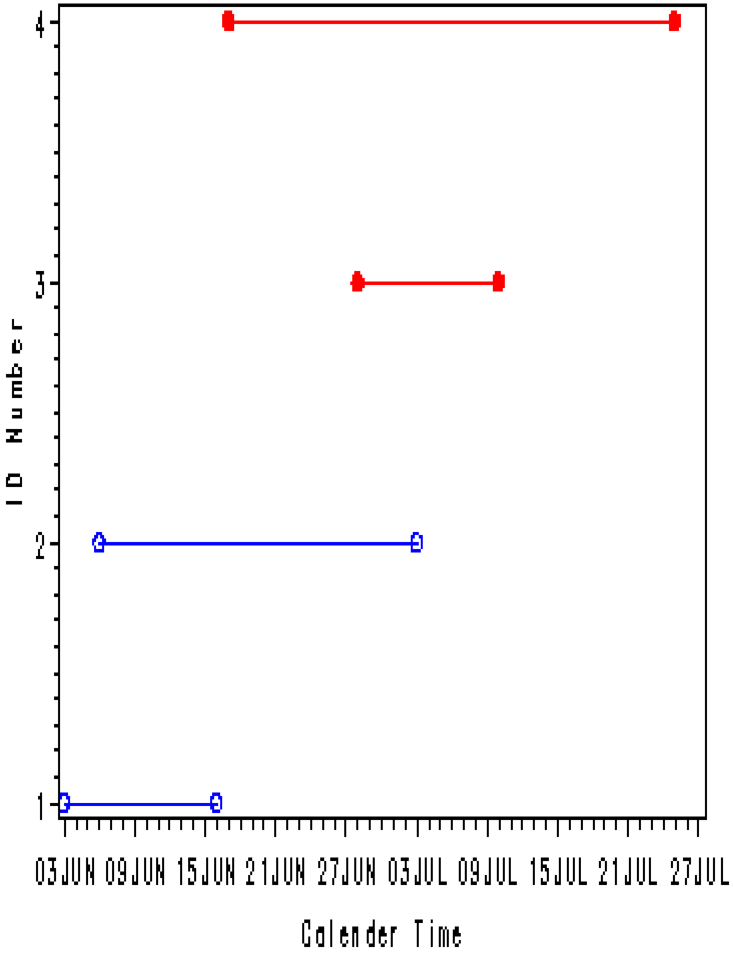
Introduction - Survival Analysis and Calendar Time

- It is important to understand the difference between calendar time and time in the study.
- It is very common for subjects to enter the study continuously throughout the length of the study.
  - For example humans may enter a mortality study when they first enter an outpatient clinic
- This situation is reflected in the graph where we can see the staggered entry of four subjects.



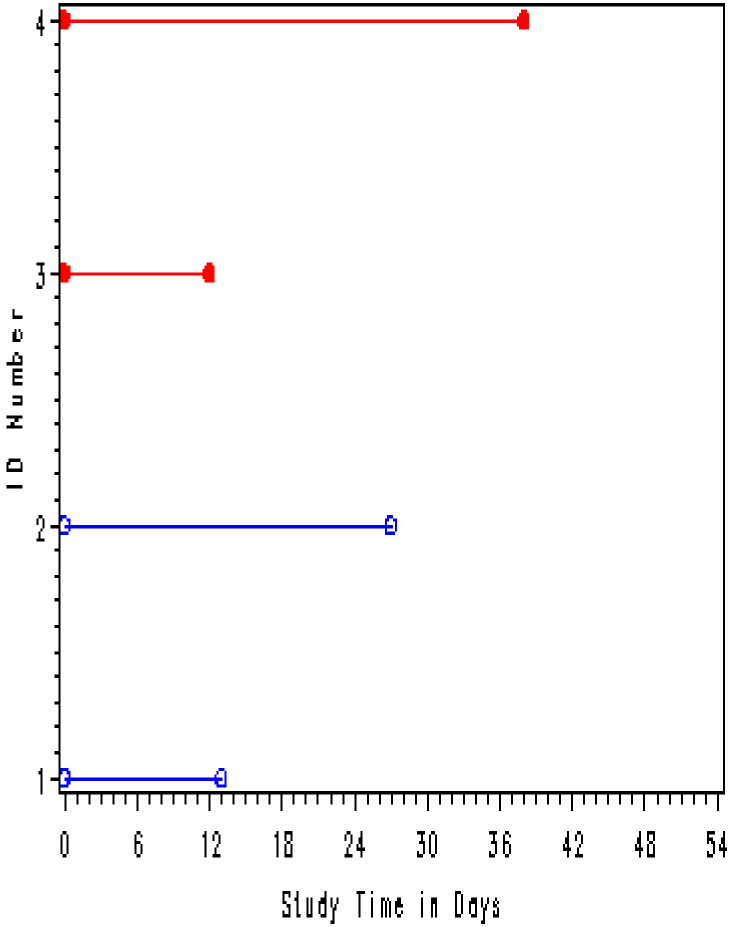
Introduction – Censoring and Calendar Time

- The subjects in red were censored and the subjects in blue experienced an event
- It would appear that
  - Subject 3 dropped out after only a short time (moved to another country)
  - Subject 4 did not experience an event by the time the study ended
    - If the study had gone on longer we would have know the time when this subject would have experienced an event.
- Thus, in calendar time both the entry and the exit time of the subjects are staggered and can occur at any time throughout the course of the study.



Introduction - What is Study Time?

- So we define study time
  - Which is the length of calendar time that the subjects were a part of the study.
- Thus,
  - Every subject starts at study time zero
  - They have ending points corresponding to the entire length of time that they participated in the study, i.e.
    - Until they experienced an event or were censored.



Further impacts of censoring

- Suppose we have a medical experiment which focuses on 8 individuals
- At the end of the observation period
  - individuals 1, 4, 5 and 8 die (D) during the course of the study
  - individuals 2 and 7 are lost to follow-up (L),
  - and individuals 3 and 6 are still alive (A).
- As far as each patient is concerned, the trial begins at some time *t0.*
- The period of time that a patient spends in the study, measured from that patient's time origin, is often referred to as *patient time.*
- *The period of time from the time origin to* the death of a patient (D) is then the survival time, and this is recorded for individuals 1, 4, 5 and 8. The survival times of the remaining individuals are right-censored (C).

Further impacts of censoring (cont.)

- In practice, the actual data recorded will be the date on which each individual enters the study, and the date on which each individual dies or was last known to be alive.
- The survival time in days, weeks or months, whichever is the most appropriate, can then be calculated.
- When survival data are to be analyzed at a predetermined point in calendar time, or at a fixed interval of time after the time origin for each patient,
  - The prognosis for individuals who are still alive can be taken to be independent of the censoring
    - Assuming that the time of analysis is specified before the data are examined.

What is the Hazard Rate?

- The other important concept in survival analysis is the hazard rate.
- From looking at data with discrete time (time measured in large intervals such as month, years or even decades) we can get an intuitive idea of the hazard rate.
- For discrete time the hazard rate is the probability that an individual will experience an event at time t while that individual is at risk for having an event.
  - e.g. What is the probability a human will experience a heart attack at age 60 (given they have survived to this age)
- Thus, the hazard rate is really just the (unobserved) rate at which events occur,
  - but it controls both the occurrence and the timing of the events
- *So:*
  - *if the hazard rate is constant over time and it was equal to 1.5, for example, this would mean that one would expect 1.5 events to occur in a time interval that is one unit long*
  - *Furthermore, if a person had a hazard rate of 1.2 at time t and a second person had a hazard rate of 2.4 at time t then it would be correct to say that the second person's risk of an event would be two times greater at time t..*

A more formal definition of the Hazard Function

- The **hazard function**, denoted by λ(t) is defined as the event rate at time *t* conditional on survival until time *t* or later (that is, *T* ≥ *t*),

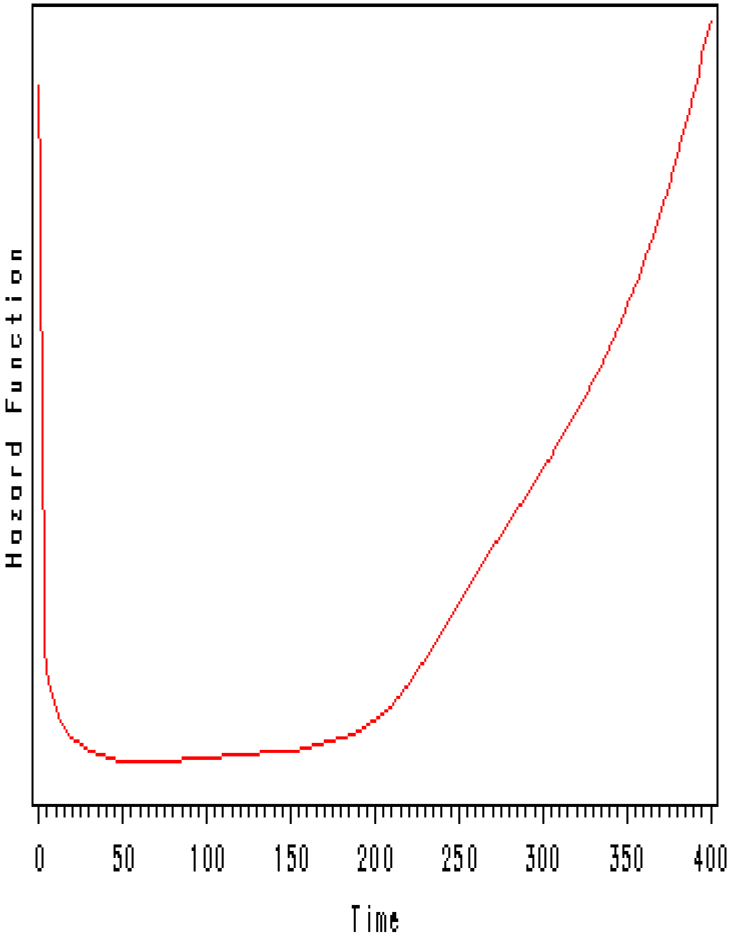
$$\lambda(t) = \lim_{dt \to 0} \frac{\Pr(t \le T < t + dt \,|\, T \ge t)}{dt} = \frac{f(t)}{S(t)} = -\frac{S'(t)}{S(t)}.$$

- The hazard function must be non-negative, $\lambda(t) \ge 0$,;

- The hazard function is widely used to express the risk or hazard of death at some time t,
  - It is obtained from the probability that an individual dies at time t, <u>but conditional on he or she having survived to that time</u>.
- It may be increasing or decreasing, non-monotonic, or discontinuous.
- Most survival analysis is based on the hazard function

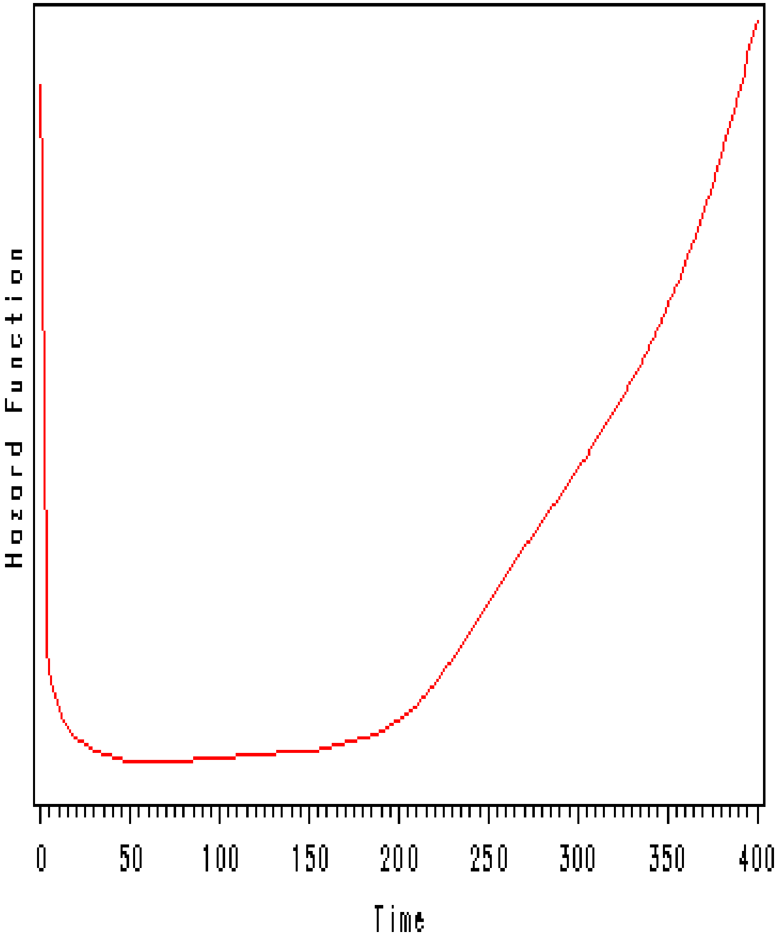
Hazard Function Example – Bath Tub Shape

- An example is the bathtub curve hazard function, which is large for small values of *t*, decreasing to some minimum, and thereafter increasing again;
- This graph is depicting the hazard function for the survival of organ transplant patients (plotted when organ transplants were in their infancy).
- At time equal to zero they are having the transplant and since this is a very dangerous operation they have a very high hazard (a great chance of dying).
- The first 10 days after the operation are also very dangerous with a high chance of the patient dying but the danger is less than during the actual operation and hence the hazard is decrease during this period.



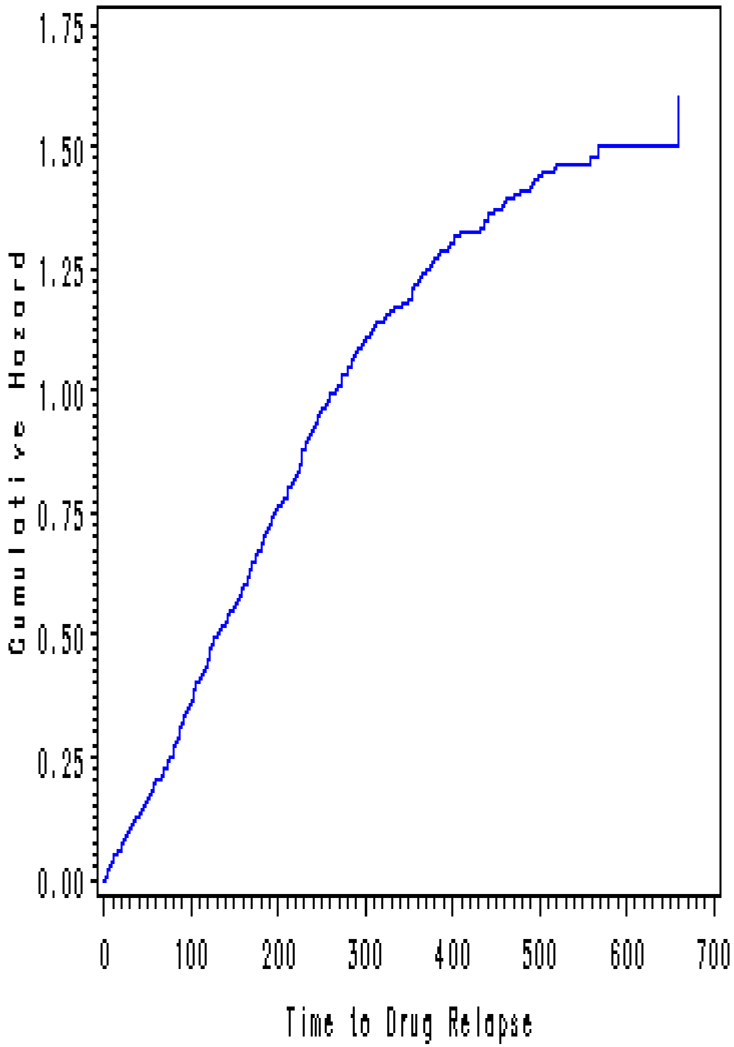
Hazard Function Example – Bath Tub Shape (cont.)

- If the patient has survived past day 10 then they are in very good shape and have a very little chance of dying in the following 6 months.
- After 6 months the patients begin to experience deterioration and the chances of dying increase again and therefore the hazard function starts to increase.
- After one year almost all patients are dead and hence the very high hazard function which will continue to increase.



Uses of the Hazard Function

- The hazard function may not seem like an exciting variable to model but other indicators of interest, such as the survival function, are derived from the hazard rate.
- Once we have modeled the hazard rate we can easily obtain these other functions of interest.
- To summarize, it is important to understand the concept of the hazard function and to understand the shape of the hazard function.
- Usually it is difficult to generate the hazard function instead we usually look at the cumulative hazard curve.

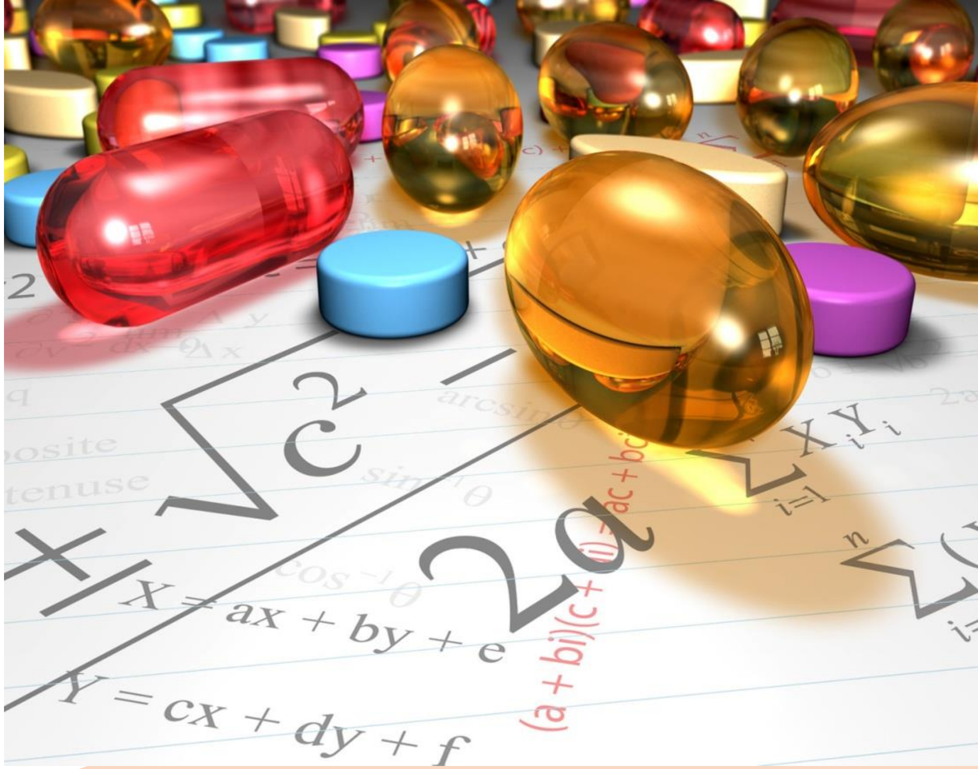


Analyzing the Hazard Function

- Another reason for modelling the hazard function is to obtain an estimate of the hazard function itself for an individual.
  - The median survival time, which will be a function of the model.
  - The median survival time can be estimated for current or future patients
- The resulting estimate could be particularly useful in devising suitable treatments regimen, or in counselling the patient about their prognosis.
- We may also be interested in looking at how different factors affect the hazard function for a particular individual or machine
- One objective of the modelling process is to determine
  - Which particular factors. or
  - which combination of potential explanatory variables

affect the form of the hazard function.

COX REGRESSION MODELS



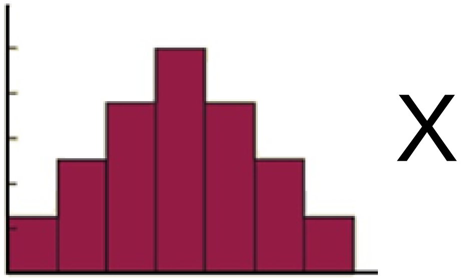
What are Cox Regression Models? When do we use these models?

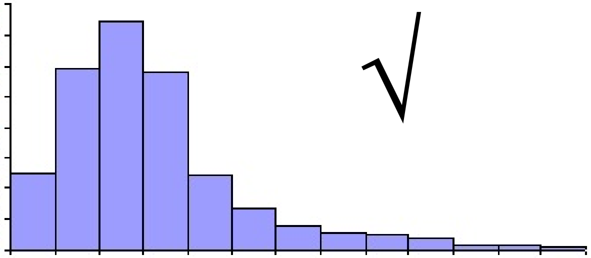
What are Cox Regression Models?

- The basic model for survival data is the *Proportional Hazards model. This model was proposed by Prof. D. Cox (1972) and has also come to be known* as the *Cox regression model.*
- *Although the model is based on the assumption of proportional* hazards, no particular form of probability distribution is assumed for the survival times.
- The model is therefore referred to as a *semi-parametric model.*
- *We now go on to develop the model for the comparison of the hazard functions for* individuals in two groups.
- It is also a *Generalized Linear Model* that we discussed in this course

What is the background to Cox Regression Models?

- When we perform survival analysis, we are usually interested in the time to the occurrence of an event of interest
  - This will be our response variable
- However survival data are not generally amenable to standard statistical procedures used in data analysis.
  - One reason is that survival data are generally not symmetrically distributed.
  - The Histogram constructed from the survival times of a group of similar individuals will tend to be *positively skewed, i.e.*
    - The histogram will have a longer “tail” to the right of the interval that contains the largest number of observations
  - Frequently this is caused by censoring





The basis of Cox Regression Models

- Suppose that patients suffering from a medical condition are randomized to receive either a standard treatment or a new treatment, and let
  - h**S(t)** is the hazards of death at time t for patients on the standard treatment
  - h**N(t)** be the hazards of death at time t for patients on the new treatment, respectively.
- According to a simple model for the survival times of the two groups of patients, the hazard at time t for a patient on the new treatment is proportional to the hazard at that same time for a patient on the standard treatment.
- This proportional hazards model can be expressed in the form

$$\mathbf{hN(t)=\psi hS(t),}$$

for any non-negative value of t, where ψ is some (multiplicative) constant.

- *NOTE: this implies that the corresponding true survivor functions for individuals on the new and standard treatments do not cross.*

Interpreting *ψ*

- The value of ψ is the ratio of the hazards of death at any time for an individual on the new treatment relative to an individual on the standard treatment, and so ψ is known as the relative hazard or hazard ratio.
- It can be interpreted as follows
  - If ψ<1, the hazard of death at t is smaller for an individual on the new drug, relative to an individual on the standard.
    - The new treatment is then an improvement on the standard.
- If ψ>1, the hazard of death at t is greater for an individual on the new drug, and the standard treatment is superior to the new treatment
- The relative hazard ψ cannot be negative, and so it is sometimes convenient to set ψ=exp(β).
  - β is also some constant ; β=log ψ, and $-\infty < \beta < \infty$
  - Note that positive values of β are obtained when the hazard ratio, ψ, is greater than unity,

Using Explanatory Variables

- Now let X_1 be an indicator variable, which takes the value zero if an individual is on the standard drug, and unity if an individual is on the new drug.
- If x_{1i} is the value of X_1 for the i^{th} individual in the study, i=1, 2,…, n, the hazard function for this individual can be written as

$$h_i(t) = e^{-\beta * x_{1i}} h_0(t)$$

  - where x_i=1 if the i^{th} individual is on the new treatment and x_i=0 otherwise.
- This is the proportional hazards model for the comparison of two treatment groups.
- Let *$h_0(t)$ be the hazard function for an* individual for whom the values of X_1 zero.
  - The function *$h_0(t)$ is called the baseline hazard function.*

Using Exploratory Variables

- To get to the Cox Regression model we now generalize to the situation where there also other explanatory variables , such as
  - Gender
  - Age,.....
  - etc…

  These are represented by variables
- Then the hazard of death at a particular time also depends on the values x_2,…, x_p of p explanatory variables, X_1, X_2,…, X_p.

  $$\text{So} \quad h_i(t) = e^{-(\beta_1 * x_{1i} + \beta_2 * x_{2i} + \,.....)}\, h_0(t)$$
- *Note:*
  - *These variables can be continuous, categorical, Binary.....*
  - *The values of these variables will be assumed to have been recorded at the time origin of the study.*

The General form of the Proportions Hazard Model

- Since this model can be re-expressed in the format,

$$\text{Log}\ (h_i(t)/h_o(t)) = \text{Log}(\psi) = \beta_1 * x_{1i} + \beta_2 * x_{2i} + \ldots\ldots$$

- The proportional hazards model may also be regarded as a linear model for the logarithm of the hazard ratio.
- It is a generalized linear model with the link function being the log function
- There are other possible forms for ψ***, but the choice*** $\boldsymbol{\psi(x_i)=exp(\beta' x_i)}$ leads to the most commonly used model for survival data.
  - Alternatives are the Weibull parametric model, the lognormal model, the Gompertz model
- Notice that there is no constant term in the linear component of the proportional hazards model.
  - If a constant term β_0*, say, were included, the baseline hazard function could simply* be rescaled by dividing $h_0(t)$ *by* $exp(\beta_0)$*, and the constant term would cancel out.*
  - *Also* we have made no assumptions concerning the actual form of the baseline hazard function $h_0(t)$*.*

Including a Variate

- Variates, either alone or in combination, are readily incorporated in a proportional hazards model.

- Each variate appears in the model with a corresponding *β-coefficient. As an illustration,* consider a situation in which the hazard function depends on two variates X_1 *and* X_2.

- *The value of these variates for the* i^{th} *individual will be* x_{1i} *and* x_{2i}*, respectively, and the* proportional hazards model for the i^{th} *of n individuals is written as* $h_i(t)=exp(\beta_1 x_{1i}+\beta_2 x_{2i})h_0(t)$.

- In models such as this, the baseline hazard function, $h_0(t)$*, is the hazard function for an* individual for whom all the variates included in the model take the value zero.

Including a Factor

- Suppose that the dependence of the hazard function on a single factor, *A, is to be modelled,* where *A has a levels.*

- *The model for an individual for whom the level of A* ^{is} *j will then* need to incorporate the term α_j *which represents the effect due to the j*th *level of the factor.*

- The terms α_1, α_2,…, α_a *are known as the main effects of the factor A.*

- *According to the* proportional hazards model, the hazard function for an individual with factor *A at level j is* exp(α_j)$h_0(t)$.

- *Now, the baseline hazard function* $h_0(t)$ *has been defined to be the hazard for an* individual with values of all explanatory variables equal to zero.
  - To be consistent with this definition, one of the α_j *must be taken to be zero.*

Including an Interaction Term

- In this situation, it may also be appropriate to include a term in the model that corresponds to individual effects for each combination of levels of two or more factors. Such effects are known as interactions.

- For example, suppose that the two factors are the sex of a patient and grade of tumor.

- If the effect of grade of tumor on the hazard of death is different in patients of each sex, we would say that there is an interaction between these two factors.

- The hazard function would then depend on the combination of levels of these two factors.

- In general, if *A and B are two factors, and the hazard of death depends on the combination* of levels of *A and B, then A and B are said to interact.*

- *Also we must (eventually) provide a physical explanation for the interaction*

An example of Proportional Hazards Modelling

- The file *Demo Cox* shows data from Kalbfleisch and Prentice (The Statistical Analysis of Failure Time Data, Wiley, 2002, p. 119)
  - It represent a clinical trial investigating the effect of covariates on time to death of patients with lung cancer.
  - We wish to determine which covariate influences the survival time
  - Within the dataset,
    - **daysurv** is the time data;
    - **censoring** is the status variable (1 for death, 0 for censored)
    - The covariates are
      - **perfstatus**
        - » The performance status of the patient at the beginning of the study
      - **Age**
        - » The age of the patient at the beginning of the study
      - Month
        - » the number of month since lung cancer diagnostic at the beginning of the study
      - **Priortherap**
        - » Whether an earlier treatment had been applied before the current trial

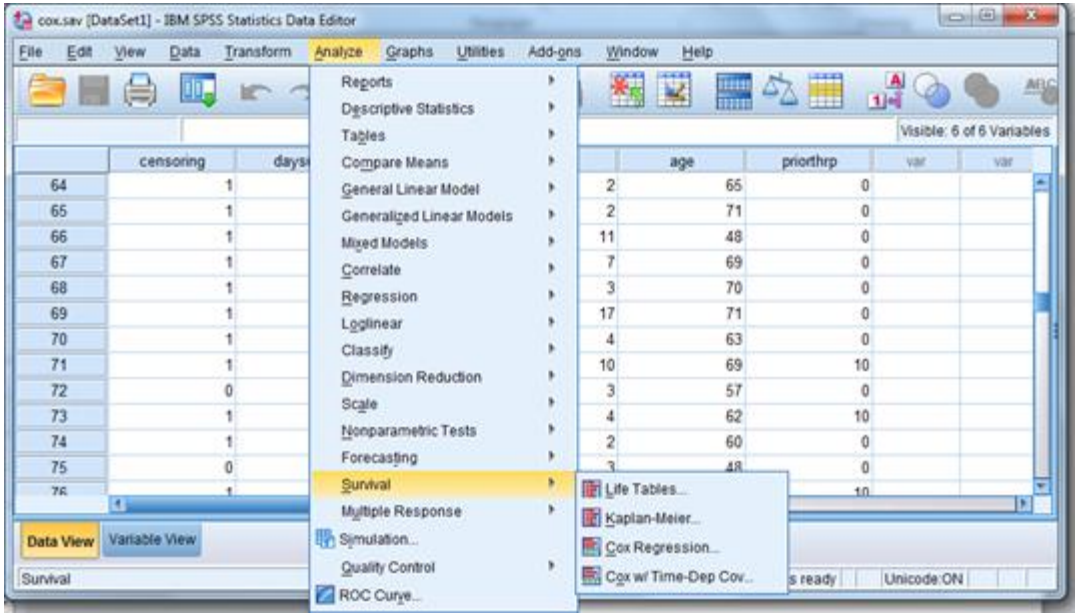
Sample of Data from *Demo Cox*

censoring	daysurv	perfstatus	month	age	priorthrp
1	1	50	7	35	0
1	1	20	21	65	10
1	2	40	36	44	10
1	3	30	3	43	0
1	4	40	2	35	0
1	7	50	7	72	0
1	7	20	11	66	0
1	7	40	4	58	0
1	8	80	2	68	0
1	8	20	19	61	10
1	8	50	5	66	0
1	8	40	58	63	10
1	10	40	23	67	10
1	10	20	5	49	0
1	11	70	11	48	10
1	12	50	4	63	10
1	12	40	12	68	10
1	13	60	4	56	0
1	13	30	2	62	0
1	15	50	13	40	10
1	15	30	5	63	0
1	16	30	4	53	10
1	18	30	4	60	0
1	18	40	5	69	10
1	18	20	15	42	0

Using SPSS to perform Proportional Hazard Modelling (I)

- After accessing and inputting the *Demo Cox* file then access

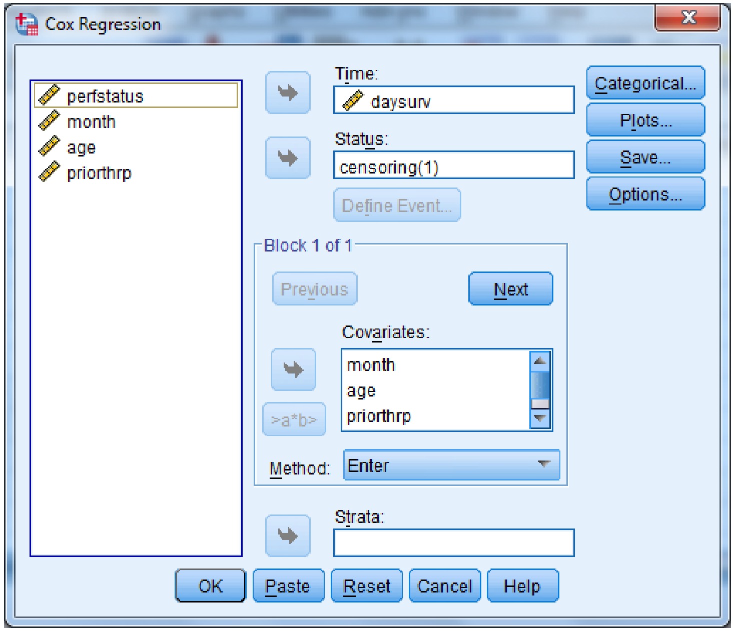
  <Analyse>, then <Survival> then <Cox Regression>



	censoring	days		age	priorthrp	var	var
64	1		2	65	0		
65	1		2	71	0		
66	1		11	48	0		
67	1		7	69	0		
68	1		3	70	0		
69	1		17	71	0		
70	1		4	63	0		
71	1		10	69	10		
72	0		3	57	0		
73	1		4	62	10		
74	1		2	60	0		
75	0		3	48	0		
76	1				10		

Using SPSS to perform Proportional Hazard Modelling (II)

- Set *daysurv* to be the (time) response variable
- *Month, age, priorthrp,* and *perfstatus* are the covariates
- *Censoring* will denote whether an event took place or the individual survived to the end of the trial



Cox Regression

Time: daysurv

Status: censoring(1)

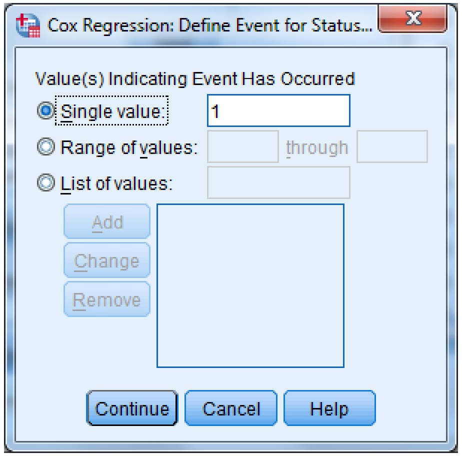
Block 1 of 1

Covariates: month, age, priorthrp

Method: Enter

Strata:

- The values are set in the <define event> dialog box



Cox Regression: Define Event for Status...

Value(s) Indicating Event Has Occurred

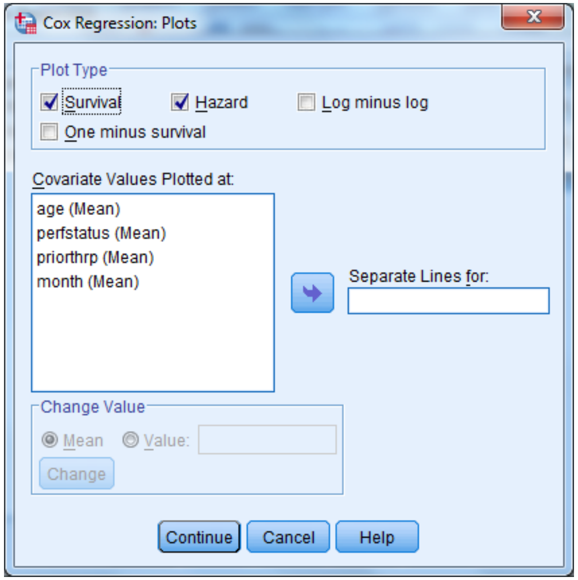
● Single value: 1

○ Range of values: through

○ List of values:

Using SPSS to perform Proportional Hazard Modelling (III)

- Select Survival and Hazard options in <Plots> dialog



Cox Regression: Plots

Plot Type

☑ Survival ☑ Hazard □ Log minus log

□ One minus survival

Covariate Values Plotted at:

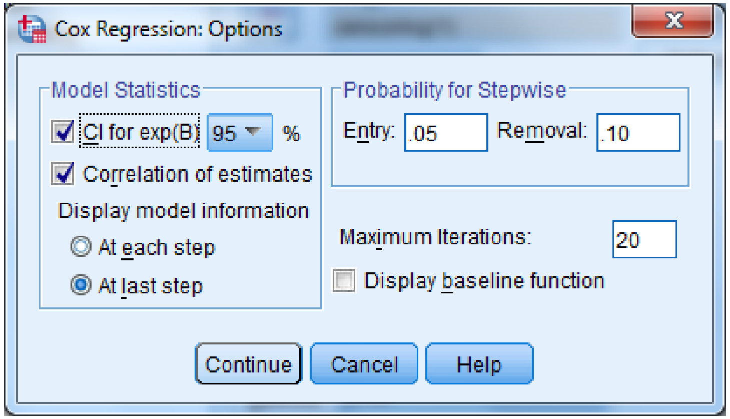
age (Mean)
perfstatus (Mean)
priorthrp (Mean)
month (Mean)

Separate Lines for:

Change Value

● Mean ○ Value:

- Ask for 95% confidence intervals and also correlation of coefficient estimates



Cox Regression: Options

Model Statistics

☑ CI for exp(B) 95 %

☑ Correlation of estimates

Display model information

○ At each step

● At last step

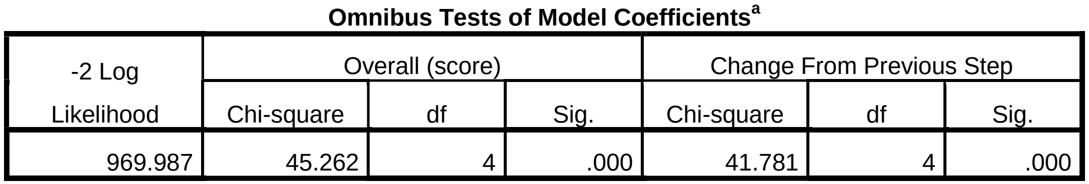
Probability for Stepwise

Entry: .05 Removal: .10

Maximum Iterations: 20

□ Display baseline function

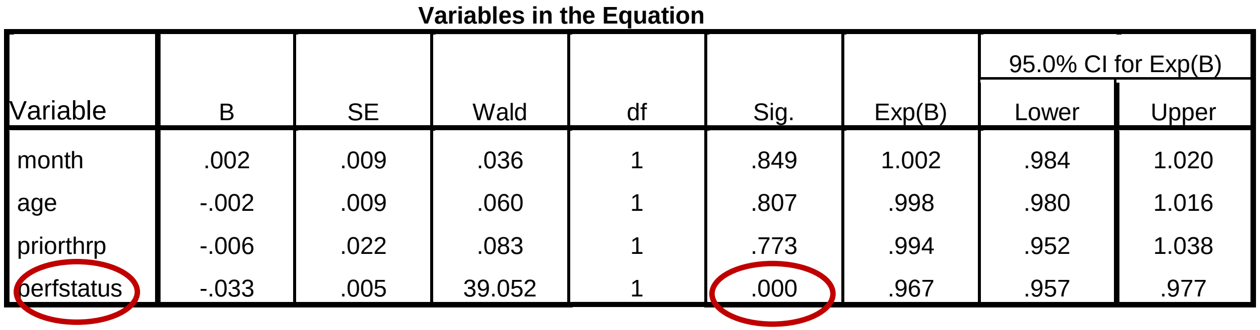
Using SPSS to perform Proportional Hazard Modelling (IV)



Omnibus Tests of Model Coefficients[a]

-2 Log Likelihood	Overall (score)			Change From Previous Step		
	Chi-square	df	Sig.	Chi-square	df	Sig.
969.987	45.262	4	.000	41.781	4	.000

- The omnibus test of model coefficients indicates that the model fitted is significant at the .001 % level

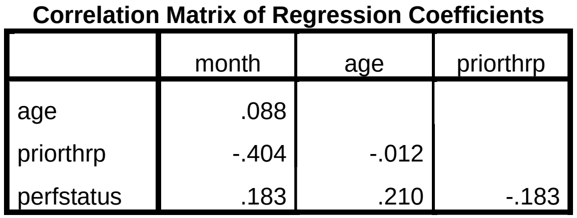


Variables in the Equation

Variable	B	SE	Wald	df	Sig.	Exp(B)	95.0% CI for Exp(B)	
							Lower	Upper
month	.002	.009	.036	1	.849	1.002	.984	1.020
age	-.002	.009	.060	1	.807	.998	.980	1.016
priorthrp	-.006	.022	.083	1	.773	.994	.952	1.038
perfstatus	-.033	.005	39.052	1	.000	.967	.957	.977

- When we look at the coefficients and their associated standard errors, we note that *perfstatus* is the only significant variable in determining *daysurv*

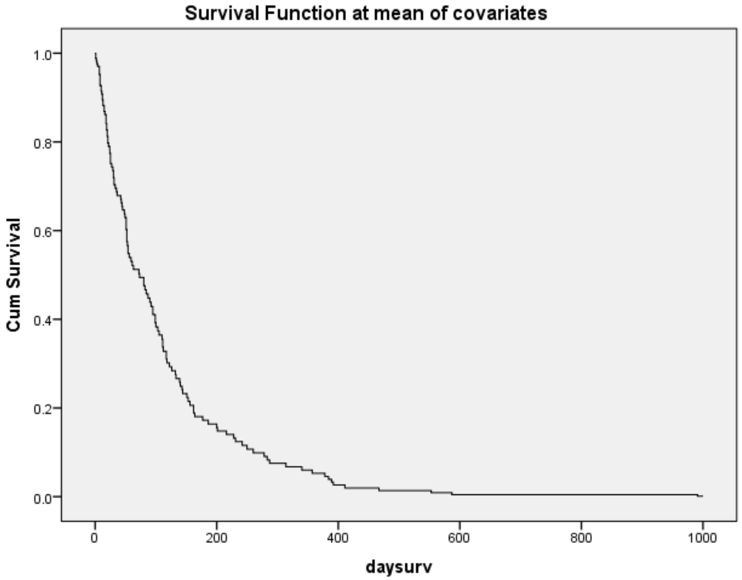
- The correlation matrix shows no strong correlation between parameter estimates



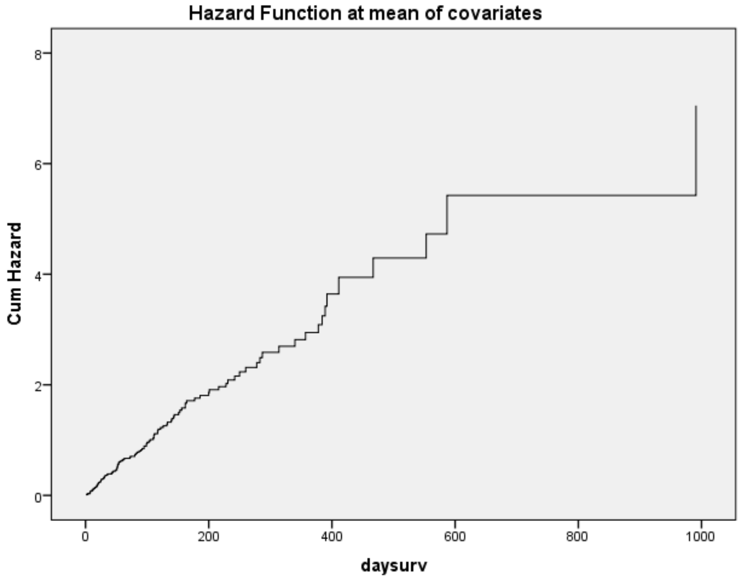
Correlation Matrix of Regression Coefficients

	month	age	priorthrp
age	.088		
priorthrp	-.404	-.012	
perfstatus	.183	.210	-.183

Using SPSS to perform Proportional Hazard Modelling (V)



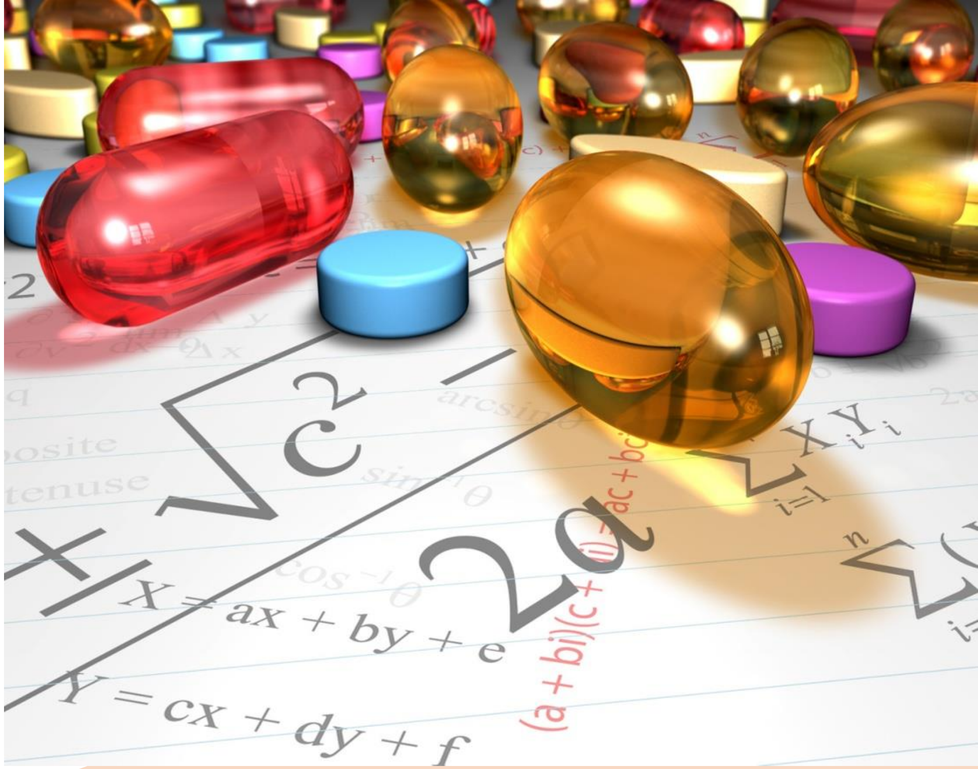
Survival Function at mean of covariates



Hazard Function at mean of covariates

- These show the variation of the survival function and the hazard function

KAPLAN MEIER CURVES



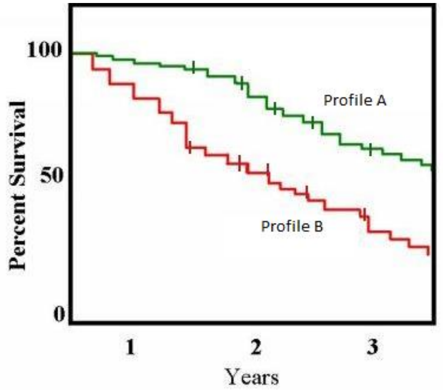
How are Kaplan Meier Curves Used?

Kaplan Meier Curves

- The **Kaplan-Meier Estimator** [Kaplan and Meier (1958), also called *the product limit* estimator, is the default estimator of the survival function
  - It is used by most software packages.
  - It is a **non-parametric** estimator, that is it incorporates no explicit explanatory variable representations
- This estimator incorporates information from all the observations available, both uncensored (event times) and censored, by considering survival to any point in time as a series of steps defined at the observed survival and censored times.
- We use the observed data to estimate the conditional probability of confirmed survival at each observed survival time and then multiply them to obtain an estimate of the overall survival function
  - It can be used with censored data
- It is widely used in the medical, engineering and insurance fields

What is the Kaplan Meir Estimator?

- A plot of the Kaplan–Meier estimator is a series of declining horizontal steps which, with a large enough sample size, approximates the true survival function for that population.
  - The value of the survival function between successive distinct sampled observations ("clicks") is assumed to be constant.
- An application in medical statistics, may involve grouping patients into categories, for instance, those with a gene A and those with gene B. In the graph, patients with Gene B die much more quickly than those with gene A.
  - After two years, about 80% of the Gene A patients survive, but less than half of patients with Gene B.
- To generate a Kaplan–Meier estimator, you need for each patient
  - The group to which the patient belongs
  - The status at last observation (either event occurrence or right-censored)
  - The time to event (or time to censoring).



Calculating the Kaplan Meir Estimator

- Let S(t) be the probability that a member from a given population will have a lifetime exceeding time, t. For a sample of size N from this population, let the observed times until death (or loss to follow-up) of the N sample members be

$$t_1 \leq t_2 \leq t_3 \, . \, ... \leq t_N$$

- Corresponding to each t_i is n_i, the number "at risk" just prior to time t_i, and d_i, the number of deaths at time t_i.
  - Note that the intervals between events are typically not uniform..
- The Kaplan–Meier estimator is the nonparametric maximum likelihood estimate of S(t), where the maximum is taken over the set of all piecewise constant survival curves with breakpoints at the event times t_i.

$$S(t) = \prod_{t_i < t} (n_i - d_i)/n_i$$

- When there is no censoring, ni is just the number of survivors just prior to time ti.
- With censoring, ni is the number of survivors minus the number of losses (censored cases).

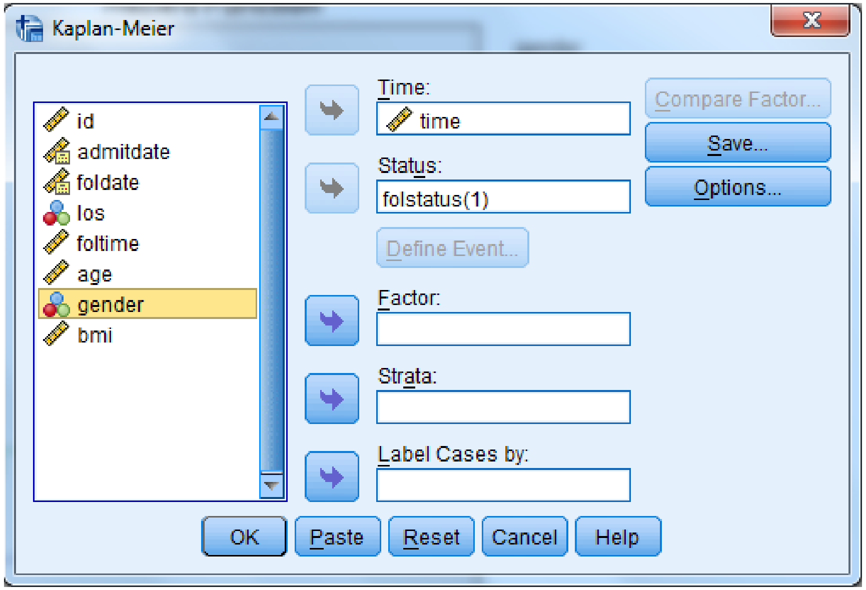
Example of Kaplan –Meier Curves (I)

- We will use the data set “**whas100.sav**“” , which contains details from the Worcester Heart Attack Study by Dr. Robert J. Goldberg of the Department of Cardiology at the University of Massachusetts Medical School
- The data describe the survival times and factors associated with patients admitted to hospitals in the Worcester Massachusetts Standard Metropolitan Statistical Area with for acute myocardial infarction
- The data contains 10 variables

Variables	Name	Description	Values/Codes
1	id	ID Code	1-100
2	admitdate	Admission Date	mm/dd/yyyy
3	foldate	Follow Up Date	mm/dd/yyyy
4	los	Length of Hospital Stay	Days
5	lenfol	Follow Up Time	Days
6	fstat	Follow Up Status	1 = Dead, 0 = Alive
7	age	Age	years
8	gender	Gender	0 = Male, 1= Female
9	bmi	Body Mass Index	kg/m^2
10	Time	los/365.2 (Length of Hospital Stay in years)	years

Example of Kaplan –Meier Curves (II)

- After accessing the *Kaplan-Meier* option in the *survival* menu option in *Analyse* , the following screen appears
- Select *time* as the Time variable, *folstatus* as the Status variable

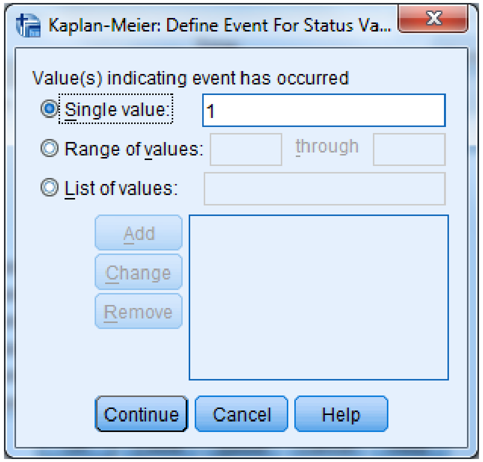


Kaplan-Meier

Time: time

Status: folstatus(1)

- We define the Event as 1 (i.e. if the subject has died)



Kaplan-Meier: Define Event For Status Va...

Value(s) indicating event has occurred

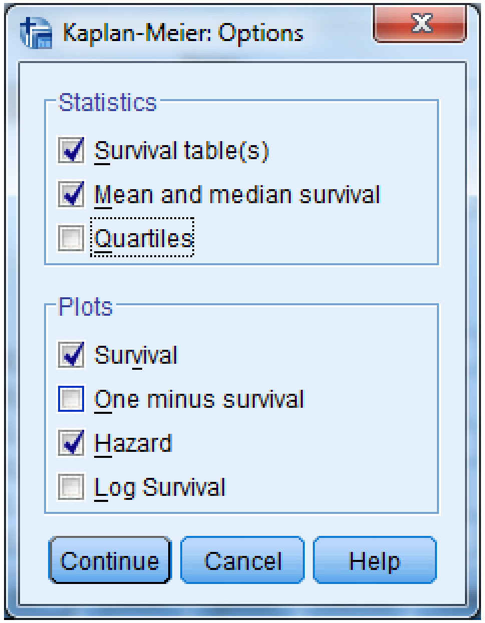
● Single value: 1

○ Range of values: through

○ List of values:

Example of Kaplan –Meier Curves (III)

- In the *options* pane we can select the *survival tables* and *mean and median survival* values.
- Also we would like to see *survival* plots and *hazard* plots



Kaplan-Meier: Options

Statistics

☑ Survival table(s)

☑ Mean and median survival

☐ Quartiles

Plots

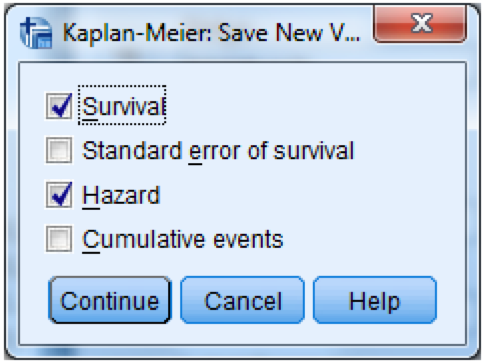
☑ Survival

☐ One minus survival

☑ Hazard

☐ Log Survival

- We can also save the *survival* function vales and *hazard* values for each subject



Kaplan-Meier: Save New V...

☑ Survival

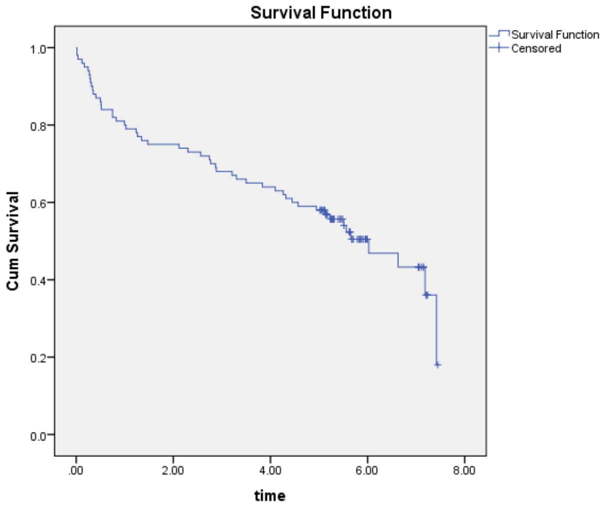
☐ Standard error of survival

☑ Hazard

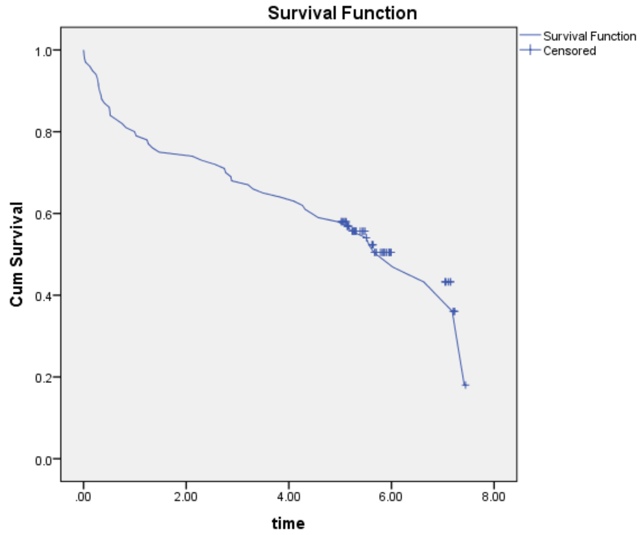
☐ Cumulative events

Example of Kaplan –Meier Curves (IV)

- The calculated survival function is shown here

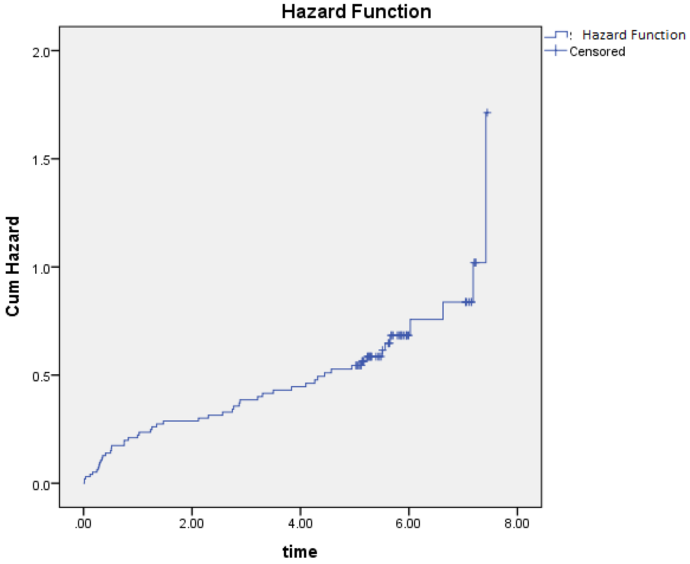


- By editing the graph we can make this a continuous function as opposed too a discrete step function



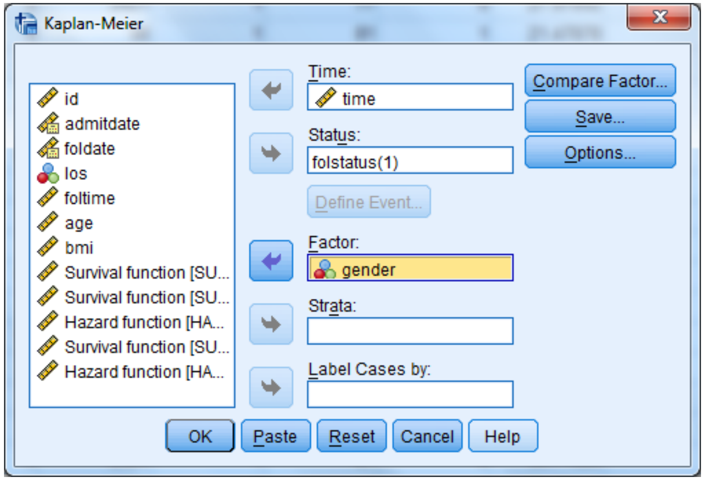
Example of Kaplan –Meier Curves (V)

- The calculated hazard function is shown here



Example of Kaplan –Meier Curves (VI)

- We can look at different survival curves for different groups
- Here we select *gender* as a factor


Kaplan-Meier

Time: time

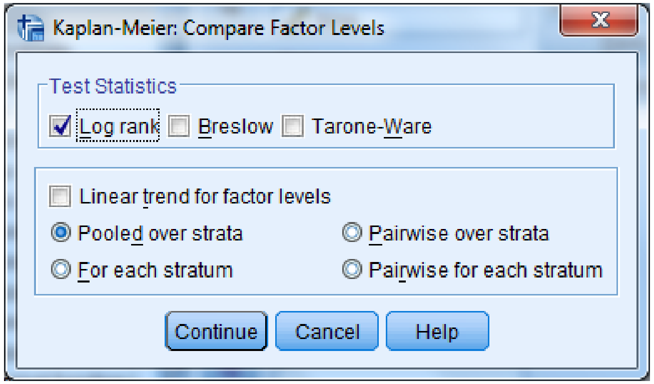
Status: folstatus(1)

Factor: gender

Strata:

Label Cases by:

- We also select the log rank test to see if there is a significant difference between the survival curves for male and female


Kaplan-Meier: Compare Factor Levels

Test Statistics

☑ Log rank ☐ Breslow ☐ Tarone-Ware

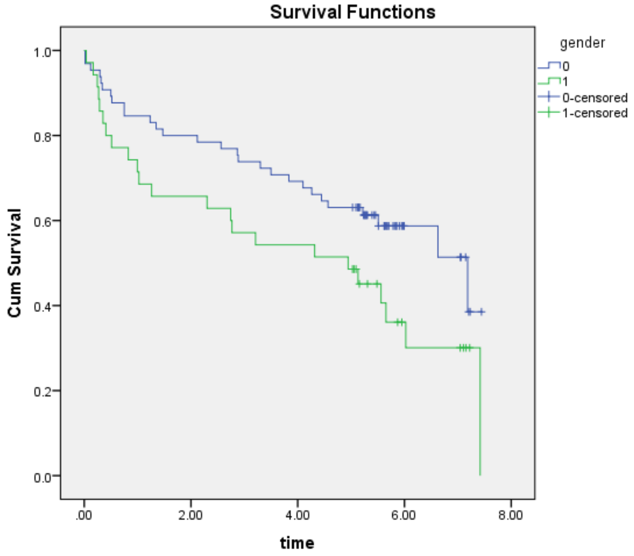
☐ Linear trend for factor levels

● Pooled over strata ○ Pairwise over strata

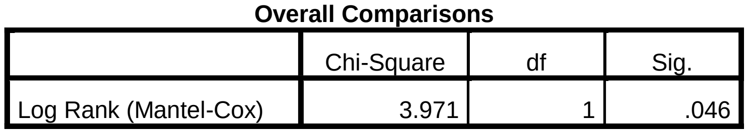
○ For each stratum ○ Pairwise for each stratum

Example of Kaplan –Meier Curves (VII)

- Clearly there is a difference in the survival curves due to gender



- The log rank test shows there is a significant difference (< 5%) between the survival curves for male and female



Overall Comparisons

	Chi-Square	df	Sig.
Log Rank (Mantel-Cox)	3.971	1	.046

References

1. Introduction to SAS. UCLA: Statistical Consulting Group. from http://www.ats.ucla.edu/stat/sas/notes2/ (accessed November 24, 2007).
2. *www.youtube.com.sg/watch?v=wTLsw-Ckfvw*
3. *www.youtube.com/watch?v=nSay4TW65dw*
4. "Applied Survival Analysis – Regression Modeling of Time-to-Event Data" by Hosmer, Lemeshow and May, 2008, Wiley
5. "Analysis of Survival data" by Cox and Oakes 1984, Chapman and Hall
6. http://www.biostat.jhsph.edu/courses/bio624/datasets/datasets.htm - datasets