inspection for

DSB - 3 Date - 24/06/22

✓ Survival Analysis

Survival Analysis is a collection of statistical procedures for ~~better~~ data analysis for which the outcome variable of interest is time untill an event occures, By time we mean years, months, weeks or days from the begining of follow up of an individual until and event occures, alternatively time can reffer to the age of an individual when and event occurs. By event we mean death disease incedence, relapse from remission, recovery or any designated experience of interest that may happen to an individual.

Censoring

Most survival analysis must consider a key analytical problem called censoring. Censoring occures when we have some information about individual survival time, but we do not know the survival time exactly.

Example:- As a simple example consider leukemia patient followed until they go out of remission. If for a given patient the study ends while the patie still in remission. (ie. doesn't get the event) that patient survival time is consider censored. For this person time is atlest as long as the period that the person has been followed, but if the person goes out of remision after the study ends, we don't know the comple survival time.

(doesn't get the event → remission)

There are generally three reason for censoring. They are —

① A person doesn't experience the event before the study ends ;

② A person is lost to follow up during the study period .

③ A person withdraws from the study .

DSE-3 [illegible] Date - 01/07/22

- **Types of Cencening :-**

There are three types of Conceening :

① **Right censoned** :

True survival time is equal to on gneaten then obsemed survival ~~nate~~ time.

② **Left censoned** :

True survival time is less then on equal to the obsenved survival time.

③ **Intenval censoned** :

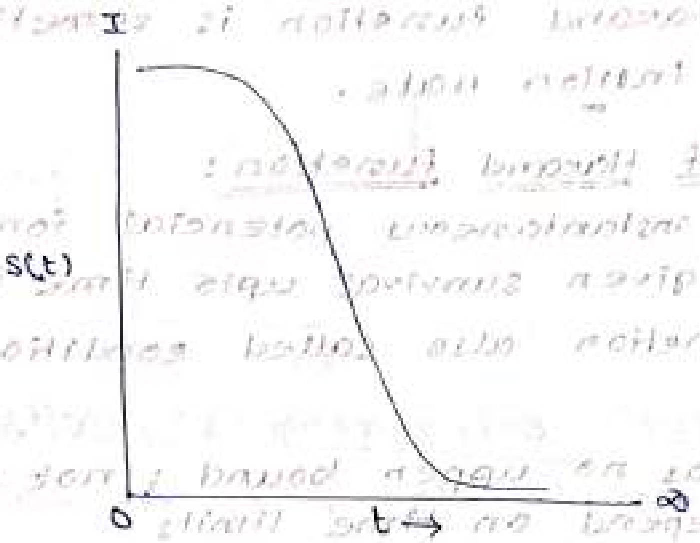
True survival time is with in a know time survival.

8

Surviven function :

The surviven function S(t) this the probability that a person survive longen then some specifiee time t i.e S(t) gives the probability that the nandom vaniable T exists the specified time t.

Theonytically as t nanges from 0 to infinite the survivon function can be natte as a smooth curve As clastnated by the group whene t identyfies the x exis all survival function have the following chanecteristic :-



① They ane nonincneasing, i.e. they head down wond as t incneases.

② At time t=0 , S(t) = S(0) = 1, i.e. at the stant of the study since know one can gotten have even the probability of surviving past time 0 is one

③ At time t=∞ , S(t) = S(∞) = 0 , i.e. theonytically if the study peniod incnease without limit eventially nobody would survivoun. so the survival past fall to 0.

Hazand function :-

The Hazand function denoted by

$$h(t) = \int \lim_{t \to 0} / \Delta t$$

The hazand function h(t) gives the instantaneous potensial pen unit time fon the even to occun, given that the individual has survive upto time t. In contnas to the survival function which foccuse on <u>not</u> failing on hazand function focus on falling i.e. on the event occuning. these in some sence, the hazand function can be considened as given the opposite ~~time~~ site. of the information, giving by the survival function.

In the hazand formula, the conditional probability gives the probability that a penson survival time T WE lime. the time interval between T an t+δt, given that survival time is greaten then on equal to t. Because of the 'given' sign ~~mean~~ here the hazand function is sometimes conditional failon nate.

- **Property of Hazand function:**

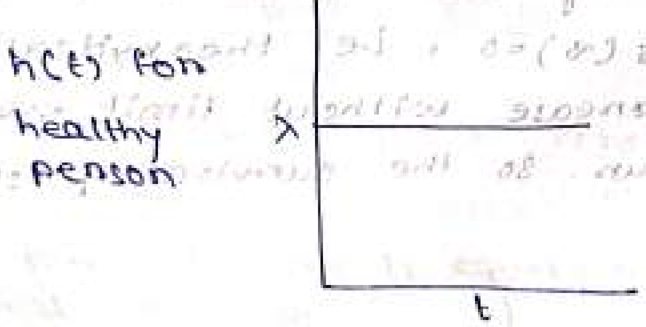
① H(t) gives instantaneous potencial fon even to occur to given survival upto time t.

② Hazand function also called conditional failon nate.

③ H(t) ≥ 0 ; has no uppen bound ; not a prob ; depend on time limits

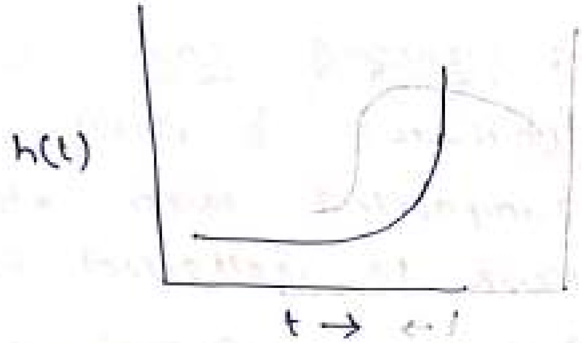
Types of Hazand function :- ✓

① **Hazand function is constant** : The following graph given shows constant hazand fon study a healthy penson.



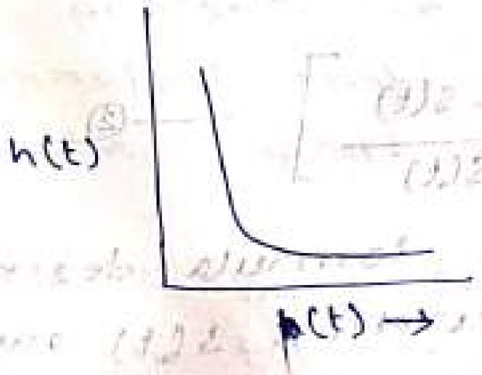
In this graph no matten what value of t specified h(t) = the some value, Hene h(t) = λ. when the hazand function is constant we say the survival model is exponential.

② **Hazand function In incneasing ovenlime :** The following graph shows a hazand function ie. increasing overtime. And example of this kind of graph is called and increasing weibull model. Such a graph might be expected for leukemia patients not responding to treatment where the event of interest is death. As survival time increases for such a patient and as the prognosis accordingly worsens, the patients potential for ~~dieg~~ dying of the disease also increases.



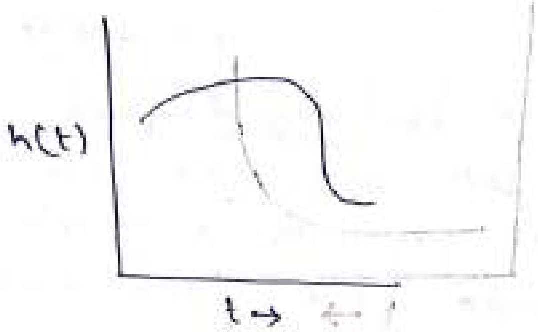
③ Hazard Function is decreasing Overtime :

In the following graph the hazard function is decreasing overtime. An example of this kind of graph is called (a) decreasing weibull. Such a graph might be expected when the event is death in persons who are recovering from surgery, because the potential for dying after surgery usually decrises as the time after surgery increases.



(4) Hazard function i.e. first increasing and then decreasing :

The following graph shows a hazard function i.e. first increasing and then decreasing. An example of this type of graph is the log normal survival model. We can expect such a graph for Tuberculosis Patients, since there potential for dyeing increases early in the disease and decreases later.



Relationship of ~~S.T. S(t)~~ S(t) and H(t) :-

There is a clearly defined relationship between the survival function (S(t)) and Hazard function (H(t)). In fact if one knows the form of S(t), one can derive the corresponding H(t) and the viceversa. The relationship between S(t) and H(t) can be expressed equivalently in either of two calculus formula shown in following:

$$S(t) = e^{-\int_0^t h(u)\,d(u)} \quad \text{—} \; ①$$

$$H(t) = -\left[\frac{\frac{d}{dt} S(t)}{S(t)} \right] \quad \text{—} \; ②$$

The first of this formula describes how the survivoun function S(t) can be written in terms of an intregral involving the Hezard Function. The formula says that s(t) equals the exponential of the -ve integral of the Hazard Function betwe integration units of 0 and t. ~~Th~~

The second formula describes how the Hazard Function H(t) can be written in terms of derivative involving the survivour function. These formula says that H(t) equals − the derivative of s(t) with respected to t devided by s(t).

Goals of Survival Analysis :-

The basic goals of survival analysis are -

① To Estimate and interpret survivour and / or hazard function of survival data.

② To compare survivour and / or hazard functio

③ To esses the relationship of explanatory variab to survival time.

Kaplan-Meier Survival Curves :-

To estimate the survival probability at a given time we make use of the risk set at that time to include the information we have on a censored person upto the time censorship, rather than simply thorough a away all the information on a censored person. The exctual compotatio of such survival prob. can be carried out using the Kaplan-Meier method (KM method).

The general formula for a KM survival prob. at failure time $t_{(f)}$ ~~at~~ is given by

$$\hat{S}(t_{(f)}) = \hat{S}(t_{(f-1)}) \times \hat{P}_r(T > t_{(f)})/(T \geq t_{(f)})$$

This formula gives the prob. of surviving past the privious failure time $t_{(f-1)}$, multiplied by the conditional probability of surviving past time $t_{(f)}$, given survival to atleast time $t_{(f)}$. The above KM formula can also be expresed as a product limit if we substitute for the survival probability $\hat{S}(t_{(f-1)})$, the product of all fraction that estimate the conditional prob. for failure times $t_{(f-1)}$ and earlier.

i.e. $\hat{S}(t_{(f-1)}) = 1 \prod_{i=f} \left[\hat{P}_r (\hat{T} > t_{(i)} / T_i \geq t_{(i)} \right]$

date - 12/07

① The remission times (weeks) for two groups of lucomia patients :

Group 1 (n=21)	Group 2 (n=21)
Treatment	Placebo
6, 6, 6, 6, 7, 10, 13, 16, 22, 23, 6+, 9+, 10+, 11+, 17+, 19+, 20+, 25+, 32+, 32+, 34+, 35+	1, 1, 2, 2, 3, 4, 4, 5, 5, 8, 8, 8, 8, 11, 11, 12, 12, 15, 17, 22, 23

$t_{(f)}$	n_f	m_f	q_f	$\hat{S}(t_f)$
0	21	0	0	1
6	21	3	1	1 × 18/21 = 0.8571
7	17	1	1	0.8571 × 16/17 = 0.8067
10	15	1	2	0.8067 × 14/15 = 0.7529
13	12	1	0	0.729 × 11/12 = 0.6901
16	11	1	3	0.6901 × 10/11 = 0.6274
22	7	1	0	0.6274 × 6/7 = 0.5371
23	6	1	5	0.5371 × 5/6 = 0.4482
>23				

$t_{(f)}$	n_f	m_f	q_f	$\hat{S}(t_f)$
0	21	0	0	⟶ 1
1	21	2	0	⟶ $1 \times 19/21 =$
2	19	2	0	⟶ $19/21 \times 17/19 = \frac{17}{21}$
3	17	1	0	⟶ $17/21 \times 16/17 = \frac{16}{21}$
4	16	2	0	⟶ $16/21 \times 14/16 = \frac{14}{21}$
5	14	2	0	⟶ $14/21 \times 12/14 = \frac{12}{21}$
8	12	4	0	⟶ $12/21 \times 8/12 = \frac{8}{21}$
11	8	2	0	⟶ $8/21 \times 6/8 = 6/21$
12	6	2	0	⟶ $6/21 \times 4/6 = 4/21$
15	4	1	0	⟶ $4/21 \times 3/4 = 3/21$
17	3	1	0	⟶ $3/21 \times 2/3 = 2/21$
22	2	1	0	⟶ $2/21 \times 1/2 = 1/21$
23	1	~~0~~ 1	0	⟶ [redacted] = 0

Unit-4

Log-Rank Test :-

The Log-Rank test is a large sample χ^2-test that uses as its test criterian a statistic that provides and overall comparison of the K-M curves being compared. This statistic like many other statistic used in other kinds of χ^2-test, makes use of observed V/s expected, cell counts over chatagories of outcomes. The catagories for the Log-Rank statistics are defined by each of the ordered failure times for the entire set of data being analysed. Here, for each order failure time $t_{(f)}$, in the entire set of data, we need the nos of subjects (m_{if}) failling at that time seperatel by group (i), followed by the no. of subjects $(n_{if}$ in the risk set at that time, also seperately by group. ~~Now, expected cell counts for each group is calculated by each~~

$t_{(f)}$	m_{1f}	m_{2f}	n_{1f}	n_{2f}	q_{1f}	q_{2f}
1	0	2	21	21	0	0
2	0	2	21	19	0	0
3	0	1	21	17	0	0
4	0	2	21	16	0	0
5	0	2	21	14	0	0
6	3	0	21	12	1	0
7	1	0	17	12	1	0
8	0	4	16	12	1	0
10	1	0	15	8	1	0
11	0	2	13	8	1	0
12	0	2	12	6	0	0
13	1	0	12	4	1	0
15	0	1	11	4	0	0
16	1	0	11	3	0	0
17	0	1	10	3	3	0
22	1	1	7	2	0	0
23	1	1	6	1	0	0

Now, expected cell counts for each groups is calculated by following formula:—

$$e_{1f} = \left(\frac{n_{1f}}{n_{1f}+n_{2f}}\right) \times (m_{1f}+m_{2f})$$

$$e_{2f} = \left(\frac{n_{2f}}{n_{1f}+n_{2f}}\right) \times (m_{1f}+m_{2f})$$

When two groups are being compared the Log-Rank test statistics is form using the sum of the observed - expected counts over all failure times for one of the two groups. For the two groups case the log rank test statistic is defined by

Log-rank test statistic,

$$= \frac{(O_i - E_i)^2}{V(O_i - E_i)} \quad , \; i = 1 \text{ or } 2$$

where.

$$v(O_i - E_i) = \sum_f \frac{n_{1f}\, n_{2f}\,(m_{1f} + m_{2f})(n_{1f} + n_{2f} - m_{1f} - m_{2f})}{(n_{1f} + n_{2f})^2\,(n_{1f} + n_{2f} - 1)} \quad ; \; i = 1 \text{ or } 2$$

The null hypothesis being tested is that there is no overall difference between the 2 survival curves. Under these null hypothesis the log-rank test statistics is approximately χ^2 with one degrees of freedom. i.e. $\chi^2 \simeq \sum_{i=1}^{2} \frac{(O_i - E_i)^2}{E_i} = \frac{(O_1 - E_1)^2}{E_1} + \frac{(O_2 - E_2)^2}{E_2}$

Now, if the calculated $\chi^2 > \chi^2_{~~0.05~~\,\alpha, 1}$ we reject the null hypothesis otherwise we accept it.

Formula for the Log-Rank statistics for several groups :- Date - 15/07

The Log-Rank test can also be used to compare three or more survival curves. The null hypothesis for these more general situation is that all survival curves are the same. Here we need to calculate the variances and covariances of $O_i - E_i$. For $i = 1, 2, \ldots, ~~g~~ G$ and $f = 1, 2, \ldots, k$ where, ~~g~~ G = no. of groups and K = no. of failure times, NIF = no. at risk in the ith group and the fth and the NIF = observed no. of failures of ith group and fth order to failure time and in ith group at fth ordered failure time $= \frac{NIF}{N_{1F} + N_{2F}}$

$$\Rightarrow \quad n_f = \sum_{i=1}^{g} n_{if}$$

$$m_f = \sum_{i=1}^{g}$$

~~$O_I - E_I$~~

$$V(O_i - E_i) = \sum_{f=1}^{k}\left(\frac{n_{if}(n_f - n_{if})\, m_{if}(n_f - m_f)}{n_f^2 (n_f - 1)}\right)$$

$$\operatorname{Cov}(O_i - E_i\,,\, O_l - E_l) = \sum_{f=1}^{k}\left(\frac{-n_{if}(n_{lf}\cdot m_f\,(n_f - m_f)}{n_f^2 (n_f - 1)}\right)$$

$$d = \left(O_1 - E_1 \cdot O_2 - E_2 \cdot \cdots \cdot O_{G-1} - E_{G-1}\right)^T$$

$$V = \left((V_{il})\right) \text{ where, } V_{ii} = V(O_i - E_i) \text{ and } V_{il} = \operatorname{cov}(O_i - E_i\,,\, O_l - E_l)$$

$$i = 1, 2, \cdots, G-1$$
$$l = 1, 2, \cdots, G-1$$

Then, the log-Rank statistic is given by the matrix product formula:

Log-Rank statistic = $d' V^{-1} d$ which has approximately a χ^2-distribution with $G-1$ degrees of freedom under the null hypothesis that all G groups have common survival curve.

Alternatives to the Log-Rank Test :-

There are several alternatives to the Log-Rank Test. The Log-Rank uses $O_i - E_i = \sum m_{if} - e_{if}$

i = no. of groups,

f = f^{th} failure time

This simple some gives the same weight namely unity to each failure time when combining observed - expected failure in each group. ~~Waiting the~~

Here the test statistic is,

$$\frac{\left(\sum w(t_{(f)})\,(m_{if} - e_{if})\right)^2}{V\left(\sum w(t_{(f)})\,(m_{if} - e_{if}\right)}$$

where, $i = 1m2$

f = f^{th} failure time

$w(t_{(f)})$ = weight at f^{th} failure time

The Wilcoxon, Tarone-Wane, Pato, Flamington-Hanington. Test statistics are vaniation of the log-Rank test statistics are denived by applying different weights at the f^{th} failure time. The Wilcoxon Test weights the observed - expected scone at time $t_{(f)}$ by the no. at nisk n_f, overall groups at time $t_{(f)}$. Thus the Wilcoxon-test places more emphasise on the information at the begining of the survival curve at the no. of nisk at large allowing early failure to receive more weight then leter fellow. This type of weighting may be used to esses wheather the effect of a treatment on survival strongest in earlier phases of administration and ten to be less effective over time.

The Tarone-Wane test statistics also applies more way to the early failure times by weighting the observed-expected scone at time $t_{(f)}$ by the square root of the no at nisk $\sqrt{n_f}$.

The Pato Test weightes the effect failure time the survival estimatise $S(t_{(f)})$ calculated overall groups combine.

The Flamington-Hanington Test uses Kaplan-Meier survival estimate $\hat{S}(t)$ overall groups to calculate its weights for the f^{th} failure time, $\hat{S}(t_{(f-1)})^{p}\left[1-\hat{S}(t_{(f-1)})\right]^{q}$

The Flamington-Hanington Test allowes the most flexibility in terms of the coice of ~~ways beca~~ weights because the user provides the values of p and q. For example - $p=1$ and $q=0$ and $\omega_p = \hat{S}t_{(f-1)}$

which gives more weights for the earlier survival times when $\hat{S}(t_{(f-1)})$ is closed to 1. Howeven if $p=0$ and $q=1$ then $w_f = 1-\hat{S}t_{(f-1)}$ in which case the letus survival time receives more weight. If $p=0$ and $q=0$ then $w_f=1$ and the Flamington-Harington test reduces to Log-Rank Test.

Date - 19/07/22

Cox-Proportional Hazard Model

The Cox-Proportional Hazard model is defined by $h(t,\underset{\sim}{X}) = h_0(t)\cdot e^{\sum_{i=1}^{p}\beta_i X_i}$. These model gives and expression for the hazard at time t for an individual with a given specification of a set of exploratory variables denoted by the $\underset{\sim}{X}$. i.e. the $\underset{\sim}{X}$ represents a collection of predicted variable i.e. being model for predict an individuals hazard. The Cox-model formula says that the hazard a time 't' is the product of two quantities. The 1st of these $h_0(t)$ is called the baseline hazard function. The 2nd quantity is the exponential expression to the linear sum of $\beta_i X_i$, where, the sum is over the p explanatory X variables. And important feature of these formula, which concens the proportional hazards assumption, is that the baseline hazard is a function of t but doesn't involve the X's. In contrast the exponential expression show here involves X but doesn't involve t.

The X are called time dependent X's. The cox-model formula has the property that if all the x's are equal to zero. The formula reduces to the baseline hazard function. i.e. the exponential part of the formula become e^0, which is 1. These property of the cox-model is the reason $h_0(t)$ is called the baseline function or from a slightly different purspective the cox-model reduces to the baseline hazard when no X's are in the model. Thus $h_0(t)$ may be considerd as a starting or baseline version of a hazard function prior to considering any of the exes. Another importand property of the cox-model is that the baseline hazard $h_0(t)$ is an unspecified function. It is this property that makes the cox-model is semi-parametric model.

Why Cox-Proportional Hazard Model is popular?

A key reason for the popularity of the cox-model is that even though the baseline hazard is not specified, reasorable good estimate, regression coefficient and hazard ratios of interest and adjusted survival curve can be obtein for a wide varalty of data situations. Another way of saying this is that the Cox-PH model is a "Robust model". so that the results from using the cox model will closely approximately the results for the connect parametric model. we would preten to use a parametric model if we were sure of the connect model. Thus when in doubt the cox - model will give relaible enough results so that it is a 'safe' choice of model, and the user doesn't need to worry about wheather the wrong parametric model is choosen.

AA ✓

Date - 26/07

maximum likelihood estimation of the cox- proportional hazard model :-

The maximum likelihood estimatos of the cox-model parameters are derived by maximising a likelihood function, usually denoted by L. The maximum likelihood function is a mathematical expresion which describes the joint prob. of obtaining the data actually obsered on the subjects in the study as a funition of unknown parameters (the β's) in the model being considered. L is sometimes written ~~no~~ notationally as L(β) ~~ot~~ β where β denots the collection of unknown parameters.

The formula for the cox-model likelihood function is actually called a partial function rather than a complete likelihood function. The term partial likelihood is used because the likelihood formula considers prob. only for those subjects who fail and doesn't expleitly considered prob. for those subjects who are censored. thus the likelihood for the cox-model doesn't consider prob. for all subjects and so it is called a partial likelihood. In panticulan the pantial likelihood can be witten as the product of sevenal ,, one fon each of say k failune times. Thus at the f^{th} failune time L_f denoles the likelihood at failing at this time given survival upto this time. Thus $L = L_1 \times L_2 \times L_3 \times \cdots \times L_k$

$$= \prod_{j=1}^{k} L_j$$

Thus although the pantial likelihood focuses on subjects who failed the sunvival time information prion to censonship to use those subject who ane censord. i.e. a penson who is censoned aften the f^{th} failune time is pant of the nisk set used to compute L_f even though this penson is censond laten. Once the likelihood function fon a given model the maximization process is cannied out by taking pantial denivatives of log(L) with nespect to each panameten in the model and then solving a system of equ^n,

$$\frac{\delta \log(L)}{\delta \beta_i} = 0 \quad , \quad i = 1, 2, \ldots, p$$

This solution is cannied out using itenation. i.e. the solution is based on ~~in~~ a stepwise mannen which stants a guessed value fon the solution and then succesively modifieds the guessed value untill a solution is finally obtain.

Hazand Ratio :-

In genenal a hazand ratio is define as the hazand fon one individual devided by the hazand fon a diffenent ,, . The two individual being compared can be distinguis by the values of the set of predictenn, i.e.

the x's. We can write the hazard ratio as the estimate of $\underline{h(t, \underset{\sim}{x}^*)}$ divided by the estimate of $h(t, \underset{\sim}{x})$. where $\underset{\sim}{x}^*$ denotes the set of predictors for one individual and $\underset{\sim}{x}$ denotes the set of predictors for other individuals. Thus $\hat{HR} = \frac{\hat{h}(t, \underset{\sim}{x}^*)}{\hat{h}(t, \underset{\sim}{x})}$

<u>The meaning of the PH assumption</u> :- ✓

The PH assumption requires that the HR is constant overtime, or equivalently that the hazard for one individual is proportional to the hazard for any other individual, where the proportionality cont is independent of time. ~~We know that~~ To understand the PH assumption we consider the formula for the HR that compares two different specification $\underset{\sim}{X}^*$ and $\underset{\sim}{X}$ for the explanatory variable use in the cox-model. Thus ~~HR~~

$$\hat{HR} = \frac{\hat{h}(t, \underset{\sim}{x}^*)}{\hat{h}(t, \underset{\sim}{x})}.$$

$$= \frac{\hat{h}_0(t) \times e^{\Sigma \hat{\beta}_i \cdot x_i^*}}{\hat{h}_0(t) \times e^{\Sigma \hat{\beta}_i \cdot x_i}}$$

$$= e^{\Sigma \beta_i (x_i^* - x_i)} \quad \text{doesn't involve } t$$

$$= \hat{\theta} \text{ (say)}$$

Thus once the model is fitted and the value of $\underset{\sim}{X}^*$ and $\underset{\sim}{X}$ are specified the value of the exponential expression for the estimated hazard ratio is a cont. which doesn't depend time. If we denote this cont by $\hat{\theta}$ then we can write the hazard ratio as shown above. This is a a mathamatical expression which states the proportional Hazard assumption. 99