Survival analysis

Kort sammanfattning och beskrivning av föreläsningen

Survival analysis eller överlevnadsanalys är *the study of time to event questions* vilket handlar om att studera tide tills en specifik händelse inträffar. Survival analysis kan användas vid flera tillfällen men här är några exempel:

1. Vad är sannolikheten att en person överlever fram till pension? Eller vad är sannolikheten att en person tar ut pension i minst 10 år. Detta handlar då om **demografi och ekonomi.**
2. Vad är chansen att en cancer försvinner inom fem år efter en lyckad behandling. Detta handlar om **medicin.**
3. Vad är chansen att en dator inte behöver bytas ut inom 2 år. Detta handlar om **ekonomi och engineering.**
4. Skiljer sig överlevnadstiden mellan två grupper som får olika behandlingar signifikant? Detta handlar då om **medicin och biologi.**
5. Vad är sannolikheten att en person som släpps från ett fängelse begår ett brott inom ett år efter att den släpps. Detta handlar åter igenom om **demografi**

R paket för att genomföra överlevnadsanalys

Det finns många bra sätt att genomföra överlevnadsanalys. När det kommer till programvara rekommenderas R starkt. Men man kan även använda SAS eller SPSS för genomföringen. Hur som helst används följande paket i R för att genomföring av survival analysis:

- *survival/KMsurv* för färdiga dataset att träna med.
- *km.ci* paketet för genomföra konfidensintervall och konfidensband.
- *discSurv/SurvMisc/survminer(ggsurvplot())/surcomp.*
- *ggfortify* för survival analysis och plottning.

Survival analysis — the applications

I exemplen tidigare visades att överlevnadsanalys används i alla möjliga branscher. Men beroende på vart så har de olika namn. Bl.a:

Vart?

- Medicinska studier.
- Engineerings studies: känt som **reliability theory**
- Social studied: känt som **duration analysis**

Så vad är man intresserad av att göra?

- Visualisera datamaterialet.
- Testa hypoteser.
- Göra prediktioner.
- Sample size determination: handlar om hur många subjects som vi behöver för att se se en skillnad i överlevnad.

Viktiga termer

- **Event:** an event is something we are interested in occuring (say) death, birth, getting arrested, obtaining a salary raise, onset of a disease, or being cured of a disease.
- **Failure of an individual:** is the event that occurs for a given individual. For example the person dying, giving birth, being born, landing in prison, getting a pay increase, falling ill, becoming healthy, etc!
- **Indidvidual at risk:** by individual risk we mean the event has not yet occured may in the future.
- **Population at risk:** are the collection of individuals for which the event can occur.

The beginning of survival analysis: the survival function

The survival function is the primary keystone to survival analysis. The function described **the probablitiy that something is still alive or has not yet had the**

event at a given time. The function is defined as follows:

$$S(t) = P(T > t)$$

The survival function has the following components:

- T which is a random variable (slumpvariabel) a.k.a time to event
- Ts cumulative distribution function (CDF) is defined as $F(t) = P(T \leq t)$
- $S(t)$ is the probablity to survive longer than t.

To describe the function, we have a random variable T which represents *time to an event occuring,* in example, machine breaking down or death. Ts cumulative distribution function (CDF) is the **probablity that the event has already happened by time** t.

The survival function $S(t)$ is then defined as **the probablity that the event has NOT happened by time** t. In other words if we track something over time, a machine, $S(t)$ tells us the probablity that the machine is still running at time t.

The relationship between $S(t)$ and the CDF of T

Since the CDF of T defines the probablity that the event has already happened and $S(t)$ defined the probablity that the event has not happened. We get the following relationships between the survival function and Ts CDF:

$$S(t) = 1 - F(t)$$

The relationships between the probablity density function (PDF)

If the random variable T has a density $f(t)$ then

$$f(t) = -\frac{d}{dt}S(t)$$

this must be the case because $S(t)$ decreases as probablity “dies out” over time. Think of it as a computer cannot last foreer - the probablity of it lasting over time must therefore decrease! Then we get the following relationships

$$F(t) = \int_0^t f(u)du$$

$$S(t) = \int_t^{-\infty} f(u)du$$

Is the survival function S(t) monotonic?

Short answer: yes! But to understand why the survival function is monotonic; think of it as

> As time increases - the probablity that you survive longer than t can only go down, since you cannot survive forever. Again, take the computer example, the computer cannot possibly become more likely to survive when time passes.

Therefore we get the following definition:

$$S(t_1) \geq S(t_2), \quad \text{for } t_1 < t_2$$

However to truly understand this concept; let us take an example of 100 people after a surgery:

Time t (months)	Alive (survived indivudals)	S(t)
0	100	1.00
3	90	0.90
6	75	0.75
12	60	0.60

We start with 100 individuals at time 0. At 3 months, 90 of them are still alive, so the estimated probablity of surviving beyond 3 months is thefore 90/100=0.9. At 6 months, 75 are still alive so the probablity is now 100/75=0.75. At 12 months 60 are still alive; so the probablity of surviving beyong 12 months is now 0.60

Life table: Survival analysis before it was invented

Before survival analysis, in England, they came up with the life table. This is important in survival analysis because can keep track of how many survives at which time.

To get a an understand how how we formulate this, take the following example of individuals where time is discrete:

	indivID	time	status
1	1.00	1.00	2.00
2	2.00	2.00	2.00
3	3.00	1.00	2.00
4	4.00	3.00	2.00
5	5.00	3.00	2.00
6	6.00	2.00	2.00
7	7.00	1.00	2.00
8	8.00	4.00	2.00
9	9.00	2.00	2.00
10	10.00	5.00	2.00

x	# at start n_x	Cens. w_x	At risk r_x	Deaths d_x	Pr. death q_x	Pr. surv. x p_x	Pr. surv. $> x$ S_x
1	10	—	10	3	0.3	0.7	7/10
2	7	—	7	3	3/7	4/7	28/70
3	4	—	4	2	0.5	0.5	14/70
4	2	—	2	1	0.5	0.5	7/70
5	1	—	1	1	1	0	0

$$q_x = d_x / r_x$$
$$p_x = 1 - q_x$$

$$S_x = \prod_{\text{“}y \leq x\text{”}} p_y$$

We have our table which show the individuals ID, time of event, and status (meaning that the event has happened. In our table we can take notice that three

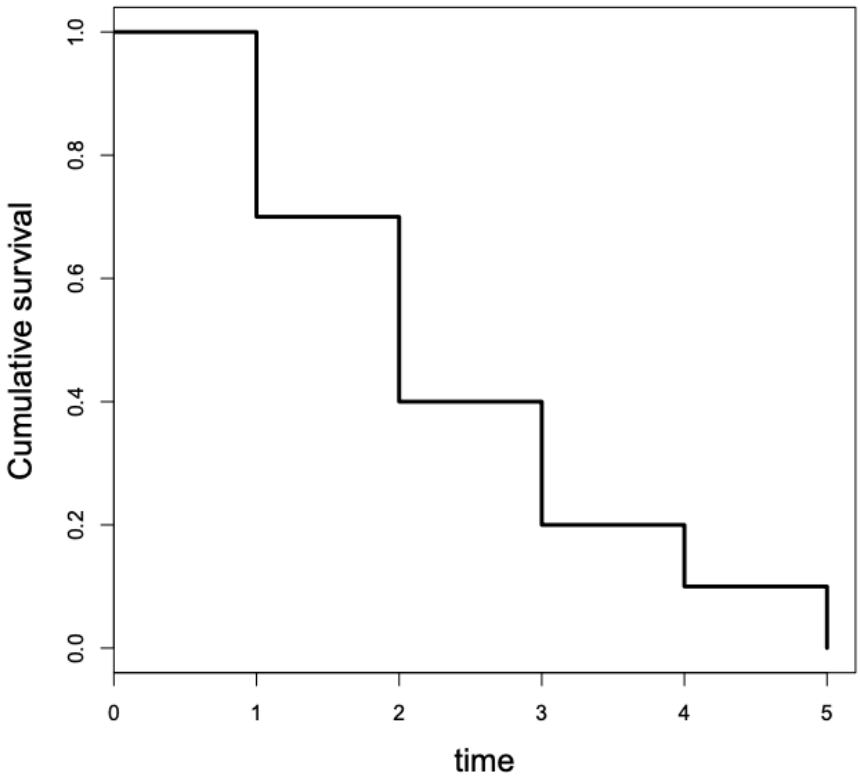
people died at time 1 (there are three 1.00), 3 died at time 2, 2 died at time 3, 1 at time 4, and 1 at time 5. These counts become the deaths column d_x

Building the life table row by row

In our table time is discrete meaning $x \in [1; 5]$. We then build our columns:

- n_x : number alive at the start of the interval x.
- w_x : censored in interval x but we have no censored data in this dataset so forget about this for now!
- r_x : "at risk" in interval x. This means the amount of people who is still alive and are at risk of dying.
- d_x. : is the amount of deaths/events in the given intervall x.
- q_x. : is the probablity of death in interval x which basically means $\frac{\text{Dead individuals}}{\text{people at risk}}$.
- p_x is the probablity of the survivors 1- probablity of death
- S_x is the probablity of survival beyond x, in example, $P(T > x)$

Plotting the survival function:



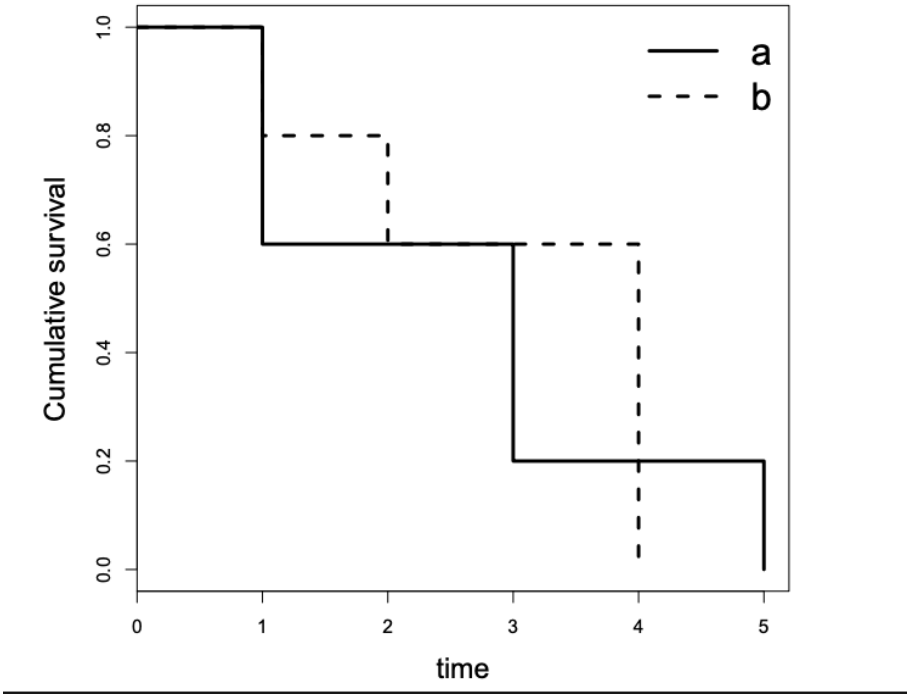
The plot of the survival function over time is called **the life curve.** In this example it's the life cruve of the 10 individuals in the dataset above. On the x-axis we plot the time and the cumulative survival which means the probablity to still be alive after time t.

We can interpret the plot as at time 0 the survival rate is 100% while at time 5 the survival rate is down to 0. For example when read $t = 4$, only 10% are expected to be alive after time = 4.

Example of survival analysis on the difference in medications a and b:

In the following table we can see individuals, time, status, and treatment:

	indivID	time	status	treatment
1	1	3	2.00	a
2	2	3	2.00	a
3	3	2	2.00	b
4	4	4	2.00	b
5	5	1	2.00	b
6	6	1	2.00	a
7	7	5	2.00	a
8	8	4	2.00	b
9	9	1	2.00	a
10	10	4	2.00	b



In this life cruve we can see that at time 0 both groups have 100% survival rate. However medication a drops to 60% survival rate at time 1 while medication b drops to 80% of surviving. We can draw the conclusion that patient in group b die earlier than group a.

Censored data

Right censoring

When we talk about right censoring we know how long the person was event-free, but never see the event time. The typical situations for this:

1. **Study ends** while the patient is still alive → we only know he survived at least until 5 years, not what happens after that.
2. **Drop-out / lost to follow up / dies from another cause →** We stop observing them at that time.

Then in the analysis we can:

> This person was at risk up to time t then their time is censored at time t.

Left censoring

When talking about left censoring we know the event already happened before we first see the person, but we dont know when it occured. Example:

At the first hospital visit the patient is already sick or already has the condition; we only know that the event happened sometime before t_0.

Interval censoring

When we hear interval censoring we dont see the exact time, but only that it occured between two observation times. Example: a patient is healthy at visit at time 1, but at the next visit at time 3 the event has already happened → event time is somewhere in (1,3) but we dont know exactly where.

Concluding on censoring

- **Right censoring:** event time is after the last time we see them (or never seen).
- **Left censoring:** event time is before the first time we see them.
- **Interval censoring:** event time is between two times we see them.

Kaplan-Meier Estimation

There is a certain problem with survival data:

- Some people **have the event** (death, relapse, etc).
- Some people are **right-censored** (study ends, drop out, die of other cause).

So we dont know the exact time event for everyone, but we still want to estimate the survival function:

$$S(t) = P(T > t)$$

Kaplan and Meier, from a 1958 paper proposed a non-parametric estimator of the survival function $S(t)$ that correctly uses both events and censoring.

The key assuption: **censoring is non-informative!**

→ Meaning that censored people are assumed to have the same future risk as those who remain under observation at that time.

How is the KM curve actually computed?

Think of the time as a sequence of events $t_1 < t_2 < ...$

At each event t_j:

- n_j : number at risk just before t_j (still in the study, not yet failed, not yet censored).
- d_j : number of events at t_j

Then we can estimate **the conditional survive through that time point:**

$$\hat{p}_j = \frac{n_j - d_j}{n_j}$$

Then the Kaplan-Meier estimate of the survival time is **the product of the conditional survivals up to** t :

$$\hat{S}(t) = \prod_{t_j \leq t} \hat{p}_j = \prod_{t_j \leq t} \frac{n_j - d_j}{n_j}$$

This also explains the reason why the curve is a STEP FUNCTION:

- It’s flat between event times.
- It drops at each even time by a factor $\frac{n_j - d_j}{n_j}$.

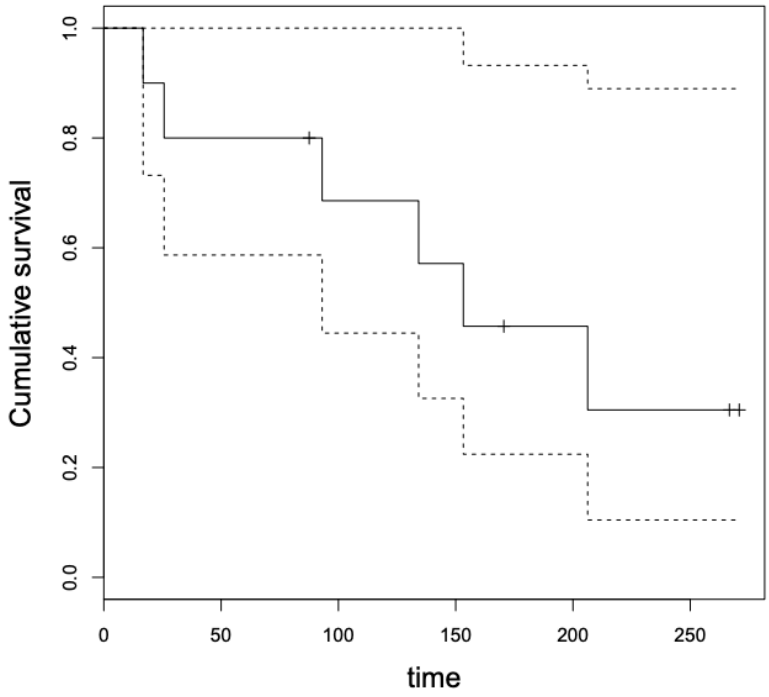
- Censored observations just reduce future risk sets n_j : they do not cause drops.

On plots, censored individuals are usually shown as small **+ sign** on the curve.

Kaplan-Meier estimation for one group

In the diagram below we can see the single group Kaplain-Meier curve. Where:

- The Kaplan Maier curve $(\hat{S}(t))$ is the black line where the + are censored individuals.
- The dotted curves around the KM curve is a 95% confidence bands for $S(t)$ often based on Greenwoods formula for the variance of $\hat{S}(t)$: we will get to this.
-



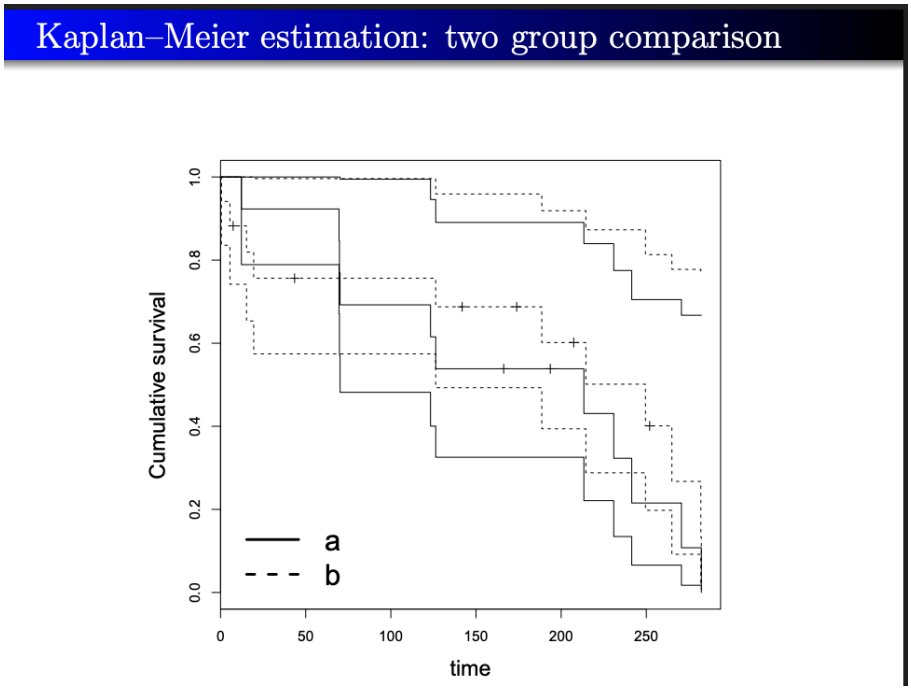
In this diagram the KM curve for the single group tells us that at each time point, the estimated probablity that a patient is still event-free (meaning the event has not yet occured), with stepts at event times, plus signs where follow-up stops, and dashed lines showing the uncertainty around that estimate.

Kaplan-Meier estimation for two group comparison

It's also possibe to calculate the KM curve between two groups. This example has group a and b. We can interpret this diagram as:

- At any time t, **the higher curve** corresponds to the group with the better survival (lower event probablity).
- If one curve is consistently above the other — this suggest that there is a treatment difference.

In theory, to test "Are the survival functions equal" we use the **log rank test** which is a paramteric test for comparing the whole curveds, and modelleing effect sizes we might use a **Cox proportional hazards models.** But visually, you already see which treatment does better at different times.



Kaplan Meier estimation: Greenwood's formula

Greenwoods formula is a way to estimate the variance (and thus the standard error) of the Kaplan-Meier survival estimator at a given time t. The formula is the standard method used to quantify the uncertainty of the Kaplan-Meier survival estimate at each time point, and it's the basis for the confidence intervals and bands you see around the KM curve.

The Kaplan-Meier estimator is

$$\hat{S}(t) = \prod_{x:x<t} (1 - q_x) = \prod_{x:x<t} (1 - \frac{d_x}{r_x})$$

here the formula consists of

1. r_x : number at risk just before time x.
2. d_x. : number of events at time x.
3. q_x. : is the estimated probablity to have the event at time x (givet you were at risk).

However just calculating the curve is not enough — we want to know how certain we it is. **Greenwoods** formula is a way to caluclate the variance of the estimator, which makes it possible to calculate the standard error. Thus we can calculate the uncertainty of the estimation (curve).

The formula is

$$v^2(t) = \hat{Var}[\hat{S}(t)] = \hat{S}(t)^2 \cdot \sum_{x:x<t} \frac{d_x}{r_x(r_x - d_x)}$$

Thus giving us

$$v(t) = \sqrt{\hat{Var}[\hat{S}(t)]}$$

Kaplan-Meier confidence intervals and confidence bands

Thanks to Greenwoods formula we are then able to calculate confidence intervals and confidence bands:

$$\text{CI} : \left[\hat{S}(t) - \Phi^{-1}\left(1 - \frac{\alpha}{2}\right) v(t),\ \hat{S}(t) + \Phi^{-1}\left(1 - \frac{\alpha}{2}\right) v(t)\right]$$

However to do this we must use the standard error and a normal approximation. This means that we need to build a CI at each fixed time t.

How to interpret

For a single chosen time point t we are about $100(1-\alpha)\%$ confident that the true survival probablity $S(t)$ likes inside that interval. In the earlier plots, there are the dashed lines around the Kaplan-Meier cure estimater.

Confidence bands VS confidence intervals

When we have pointwise confidence intervals we may sometimes get problems. Think of this as **what is the survival at a specific time?** At $t = 100$ the Kaplan Meier curve says $\hat{S}(100) = 0.7$. Using Greenwoods formula we then calculate a 95% confidence interval and we get the following reslut

$$[0.60, 0.80]$$

You can say then that at 100 days, the survival probablity is about 70%, and I am 95% confident the true value is between 60% to 80%. **This is a pointwise CI** it only talks about one chosen time point.

However there is a problem with this. The probability that all of those intervals are correct **at the same time** is less then 95% because you are check many points → multiple testing.

Change the question now!

Can we draw two curves so that the **entire true survival curve** lies between them, for all times, with about 95% probablity? This is a confidence band.

Instead of having many little intervals at seperate times, you get one shaded band (or two dashed curves) which surrounds the whole KM curve. The bands are calculated as follows:

$$\sqrt{n}(\hat{S}(t) - S(t)) \approx S(t)W(\sigma(t))$$

The problem with the confidence intervals and confidence bands

> These intervals are useless. They are good models at each time point, but not simutaneously.

What we mean by this is that pointwise confidence intervals around the Kaplan-Meier curve is the problem. At each single time t the confidence interval is fine.

But when you draw a whole bunch of these intervals along the curve and stare at them all together, you start implicity asking a **different question:**

> Is the entire survival curve well estimated/different between groups?

The intervals are not designed for this! Each confidence intervall gives you 95% coverage **only at that one time.** Ify ou look at say 50 time points, the probablity that all 50 intervals are simultaneously correct is much less then 95%. Therefore it is simply impossible to say

> The true curve is inside all of these intervals with 95% probablity.

Why does this matter when comparing two curves?

Suppose that you have Kaplan-Meiker curves for treatment A and treatment B and confidence intervals around each. A very common (but wrong) way people think is:

- "If the confidence intervals overlap, there is no difference".
- "If they don't overlap, there is a difference."

The problems with this are:

1. You are checking this at many time points → consequently doing multiple testing. At some times you'll see overlap, at other times not; you don't have a clear, global conclusion.
2. The scientific question is usually:

   > Are the two survival curves different overall? (i.e is $S_A(t) = S_B(t)$ for all t or not).

So for global questions "do the curves differ" the right tools is:

- A log rank test (non parametric test comparing whole curves) OR
- A Cox model estimating a hazard ratio with a confidence interval.

These methods consider all event times together and gives a single p-value/effect estimate.

Drawing the wrong conclusion in population comparisons

The question you must ALWAYS answer is:

> Are the populaitons comparable?

I am going to give an example in why this matters.

Say that we have a university hospital A which has a high death rate but takes on very difficult cases. Compared to other hopsital say B and C which are not taking on as difficult cases as hopsital A. Say that hospital B and C does have a high death rate compared to hopsital A. It could be the case that the wards in hospital A does much better than in B and C, however, when looking at the overall case; hospital A looks like it's doing the worst. **Therefore,** the populations are not comparable since they differ in comparison to other hospitals. This is a common mistake which many dataanalyst do not take in account for.

Risk and Odds

Before going into the Hazard Rate function and Cox regression let us talk about risks and odds.

We have the following table

	Event (hay fever)	**No event**	**Total**
Med A	$D_A = 19$	$H_A = 125$	$N_A = 144$
Med B	$D_B = 33$	$H_B = 113$	$N_A = 146$

We can now calculate the risk

The risk is the probablity of event in each group:

$$R_A = D_A/N_A = 19/144 \approx 0.132$$
$$R_B = D_B/N_B = 33/146 \approx 0.226$$

This basically tells us that

- About 13% in group A gets hay fever.
- About 23% in group B gets hay fever.

With this we can calculate the risk ratio (relative risk)

$$RR_{A/B} = R_A / R_B \approx 0.584$$

We can interpret this as the risk of hay fever in group A is about 0.58 times the risk in group B. Which means that it is rougly 42% risk in A.

Lastly we can calculate the ODDS and the odds ratio!

Then the odds in each group can be calculated as:

$$O_A = D_A / H_A = 19/125 \approx 0.152$$
$$O_B = D_B / H_B = 33/113 \approx 0.292$$

Odds ratio:

$$OR_{A/B} = O_A / O_B \approx 0.52$$

So, the odds of hay fever on group A are about half of the odds on B.

The Hazard Rate function

Let us now get into the Hazard Rate function, which can be regarded as the heart of survival analysis. **So,** we do not just care about about answering the question:

> What is the probability to survive up to time t? (That's $S(t)$)

Rather, we want to know:

> If a patient is still alive at time t. how big is the chance that they die right after t?

This is instantaneous risk, given survival so far, is the hazard. But to understand what I mean by this imagine patient in a study. At time $t = 12$ some have already had the event (died) and some a still alive. When we talk about **hazard at 12 months,** we only look at the people who are still alive **just before 12 months.** But everyone who already died is not at risk anymore. Therefore "given survival so far" just means that among those who survived up to time t.

With **instantaneous risk** we look at a very short time step after t, say from $12.0 - 12.1$. So we ask the question:

> What is the chance a person dies in this tiny interval, given that they were still alive at 12.0 months?

That chance per unit time is what we call the **hazard.** Meaning if many people die very soon after 12 months → the hazard at 12 is high, however, if almost nobody dies right after 12 months the hazard at 12 is low. **Let us take a look at the hazard function!**

The Hazard Rate (function)

The Hazard Rate function is defined as follows:

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

As you can take notice by the definiton:

> The Hazard Rate function is a derivative of something.

Inside of the probability we have the interval $[t, t + \Delta t)$. As Δt gets small, that's like looking at **how fast probablity mass is accumulating around** t**.**

1. Use the conditional probablity

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

2. Express with CDF F and survival S

$$P(t \leq T < t + \Delta t) = F(t + \Delta t) - F(t))$$
$$P(T \geq t) = S(t)$$

So this means that the hazard function goes from

$$\frac{P(t \leq T < t + \Delta t \mid T \geq t)}{\Delta t} = \frac{F(t + \Delta t)}{S(t)\Delta t}$$

If we now look at the numerator

$$\frac{F(t+\Delta t)-F(t)}{\Delta t} \rightarrow \Delta t \rightarrow 0$$

By this we mean as Δt approaches 0 we basically define the deriviate of F (CDF) at time t, i.e the density $f(t)$.

So therefore we now have:

$$\lim_{\Delta t \to 0} \frac{F(t+\Delta t)-F(t)}{\Delta t} = F'(t) = f(t)$$

Hence in the limit we have:

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

- $S'(t)$ is the deriviate of the survival function
- Divided by $S(t)$ and put a minus sign → hazard.

So when we say that this is a deriviate of something we mean that:

> The Hazard is basically **a deriviate of the CDF** (or of $-\log S(t)$ scaled by $1/S(t)$.)

What does this mean theoretically?

It means that we can say **two equivalent "deriviate views":**

- $h(t) = \frac{S'(t)}{S(t)}$ → density divided by survival.
- $h(t) = -(\log S(t))'$ → negative deriviate of log-survival.

Both $h(t)$ are derivied from the original limit.

EXPLANATION OF HOW WE WENT FROM F(t) TO S(t)

Remember the relationship between the CDF and the survival function is

$$F(t) = 1 - S(t)$$
$$F(t+\Delta t) - F(t) = 1 - S(t+\Delta t) + S(t)$$

CONCLUSION:
So in regular calculus the derivative of distance → speed (instantaneous rate of change of distance). BUT here, hazard is like **speed of failing,** but conditional on having survived to time t.

The integrated hazard

We can define the integrated (cumulative) hazard as

$$H(t) = \int_0^t h(s)ds$$

So this is just: add upp the hazard over time which → cumulative/integrated hazard.

In more detail about the integration of h(s)

The hazard $h(t)$ is the instantenous rate → "risk per unit time at eaxtly time t, given survival so far". IF we add upp this instanteous risk from time 0 to time t we get integrated hazard. We can think of the intergrated hazard as the total accumlated risk a person has been exposed to up to time t.

This means the the hazard and survival fully determine each other!

$$h(t) = -S'(t)/S(t)$$

We can rearrange this to:

$$S'(t) = -h(t)S(t)$$

Maybe you can see now that is is just an ODE; Ordinary Differential Equation. And solving this equation will give us:

$$S(t) = e^{-H(t)}$$

And since $f(t) = F'(t) = -S'(t)$ we therefore get

$$f(t) = h(t)S(t)$$

Consequently this means that if you know **Hazard** $h(t)$ you can get:

- The integrated hazard $H(t)$.
- The survival function $S(t) = e^{-H(t)}$.
- The density $f(t) = h(t)S(t)$.

So the **hazard, survival and density** are just different faces of the same thing. Specifying any one of them completely determines the others.

The Hazard, Survival, and density are different faces of the same thing

By this we mean that we have three functions that all describe the **same random time T:**

- Survival $S(t) = P(T > t)$ → What is the probablity a person is still alive after time t.
- Density $f(t)$ (if it exists) → How much probablity is concetrated around time t.
- Hazard $h(t)$ → Given a person is alive at time t, how risky is it to die right after t.

The likelihood estimation of the function

When estimating the parameters we have a survival time T, and we now assume a parametric model with the parameter θ. This model will give us:

- $f_\theta(t)$: the probablity density function (PDF)
- $F_\theta(t) = P_\theta(T \leq t)$: which is the cumulative density function (CDF)
- $S_\theta(t) = P_\theta(T > t)$ the probablity of surviving
- $h_\theta(t)$: which is the hazard.
- $H_\theta(t)$: is the cumulative hazard.

Then when want to estimate θ with maximum likelihood estimation (MLE). The important thing to understand here is that when we have estimated $\hat{\theta}$; we have automatic estimated following:

- $\hat{h}(t)$
- $\hat{S}(t)$

- $\hat{F}(t)$

So the goal is to **estimate the parameters of the survival distribution.**

The likelihood contribution for each type of observation

For each individual i; depending on what we know, their likelihood contributions are:

1. **Exact observed event at time t_i:**

   We observed the failure of time → meaning density contributes:

$$L_i(\theta) = f_\theta(t_i)$$

2. **Right-censored at time t_i**

   We only know the event happenes after t_i then the likelihood is defined as:

$$L_i(\theta) = S_\theta(t_i) = P(T \geq t_i)$$

3. **Left censored at time t_i**

   We only know the event happened before t_i then the likelihood function is defined as follows

$$L_i(\theta) = F_i(t_i) = P(T \geq t_i)$$

4. **Interval censored between $(x_i^L, x_i^R]$**

   We know the event happened in this given interval, the likelihood is then defined as follows

$$L_i(\theta) = F_\theta(x_i^R) - F_\theta(x_i^L)$$

Let us get back to theory again of what we are actually calculating in the likelihoods!

One formula that can cover all cases

In survival data each person can be:

- Event observed.
- Right-censored
- Left-censored
- Interval-censered

When calculating the MLE we could do this seperate with all four likelihoods. We could write the likelihoodfunctions as:

- Product over all events of $f_\theta(t_i)$
- Times product over all right-censored of $S_\theta(t_i)$
- Times product over all left-censored of $F_i(t_i)$
- Times product over all interval-censored of $F_\theta(x_i^R) - F_\theta(x_i^L)$

This is completely fine to do, however, it is extremely messy to do this four times. Instead we can introduce the **indicator vector** v_i and calculate all likelihoods as one single expression as follows:

$$L(\theta) = \prod_i \quad f_\theta(t_i)^{v_{i,1}} \quad S_\theta(t_i)^{v_{i,2}} \quad F_\theta(t_i)^{v_{i,3}} \quad \Big(F_\theta(x_{iR}) - F_\theta(x_{iL})\Big)^{v_{i,4}}$$

This way of calculating the MLE does have advantages:

- The formula looks the same no matter how many censored/event types you have.
- It’s easier to manipulate algebraically and to program.
- Easier to take logs, derivatives, and maximise the likelihood.

Example: exponential failure over time

Let us use an exponential model where we assume that the failure time T follows an exponential distribution with rate λ.

Consequently this means that our:

- CDF:

$$F(t) = 1 - e^{-\lambda t}$$

- Survival function

$$S(t) = e^{-\lambda t}$$

- Density:

$$f(t) = \lambda e^{-\lambda t}$$

- Hazard:

$$h(t) = \lambda \quad \text{Constant hazard}$$

- Cumulative hazard:

$$H(t) = \lambda t$$

As we can see above **the entire distribution is determined by λ.**

We now assume that we only have right censored data. Meaning:

- Data are either observed events
- or right-censored

Consequently this means that

$$v_{i,1} + v_{i,2} = 1$$

We can then formulate our log-likelihood

$$\ell(\lambda) = \sum_i v_{i,1} \log h_\lambda(t_i) - \sum_i H_\lambda(t_i)$$

Because for exponential:

- $h_\lambda(t_i) = \lambda$
- $H_\lambda(t_i) = \lambda t_i$

If we then plug in these into our formula we get

$$\ell(\lambda) = \sum_i v_{i,1} \ \log(\lambda) \ - \lambda \sum_i t_i$$

Follow define:

- $d = \sum_i v_{i,1} =$ the nuber of observed events
- $T = \sum_i t_i =$ total time at risk (sum of all survival times up to event or censoring)

And we get the following likelihood function:

$$\ell(\lambda) = d \log \lambda - \lambda T$$

Finding the MLE!

In order to find the MLE we will need to differentiate:

$$\ell'(\lambda) = \frac{d}{\lambda} - T$$

We then set it equal to zero:

$$0 = \frac{d}{\lambda} - T$$

And we solve it

$$\hat{\lambda} = \frac{T}{d}$$

This means that the maximum likelihood estimation MLE of the estimated hazard is observed event / total exposure time

> $$\text{The estimated hazard} = \frac{\text{Observed events}}{\text{Total exposure time}}$$

Making sure it's the MLE

To make sure that we have actually gotten the MLE we need to calculate the the second derivative and if it's negative we know that it's the maximum:

$$\ell''(\lambda) = -\frac{d}{\lambda^2}$$

Important insights

Remember that the information about λ depends on the number of events NOT on the number of people in the study! This is important because the more censored people without events does not equal more information. Only events contribute information about the hazard:

> Only events contribute information about the hazard. More censored people without events 0 $\neq$ more information.

Log-rank test (non-parametric)

A log-rank test is a non-parametric test which is the standard way to test if two Kaplan-Meier curves differ without assuming a specific parametric distribution.

What are we then testing?

We have two groups that we want to compare, e.g:

- Smokers vs non-smokers.
- Treatment A vs B.
- Men vs women.

Each group has its own hazard function:

- $h_1(t)$: for group 1
- $h_2(t)$: for group 2

We now look at the hazard ratio

$$HR(t) = \frac{h_1(t)}{h_2(t)}$$

We set up our hypothesis:

$$H_0 : HR(t) = 1 \text{ for all } t$$

→ this means that the two groups have the same hazard at all times → same survival curves (meaning no difference).

$$H_a : HR(t) \neq \text{for some } t$$

→ if we reject the null hypothesis the hazards differ meaning that the survival curves differ as well.

So the log-rank test check if the two KM curves are the same.

The test statistic is defined as follows:

$$\begin{aligned} Z &= \sum_x \left(d_x^{(1)} - d_x \frac{r_x^{(1)}}{r_x} \right) \\ \mathrm{Var}(Z) &= \sum_x d_x (r_x - d_x) \frac{r_x^{(1)} r_x^{(2)}}{r_x^2 (r_x - 1)} =: s^2 \end{aligned}$$

The test variable is then

$$\frac{Z^2}{s^2} \sim \chi^2$$

- If the p-value is below significance level (0.05) then survival distributions can be assumed to be different. **This is also known as Cochran-Mantel-Haenszels test.**

Cox-regression (Cox proportional hazard regression)

Cox regression is a statistical method used in survival analysis to examine the relationship between the time until and event occurs and several independent variables (covariates).

So what is Cox Regression solving?

We want to answer question like:

- Does treatment A vs B change the risk of failiure?
- How does age, smoking, blood pressure affect survival?

Hence we need a regression models where we have explanatory variable $(x_1, ..., x_k)$ and a respons variable which is the survival time (through the hazard $h(t)$).

A fully parametric model (exponential, Weibull,...) would require us to guess the exact shape of the hazard function over time which is often hard to justify. **Cox regression solves this by:** Not specifying a shape of $h(t)$ over time, but specifying how covariates act on the hazard. This is the reason why people refer to Cox regression as **semi-parametric.**

Basic theory of Cox Regression

We define the model as:

$$h(t) = h_0 \exp(\beta_1 x_1 + ... + \beta_k x_k)$$

- $h(t)$ is the hazard rate for an individual time t
- $h_0(t)$ is the baseline hazard (all $x_i = 0$), the intercept sits here.
- $\beta_1, ..., \beta_k$ are the coefficients that modify the baseline hazard.
- x_i are covariates (explanatory variables).

  With Hazard rate for an individual at time t, we mean the instanteous risk that this person will have the event right after the time t, assuming they have survived up to time t, givet their covariates.

If we set the covariates to 0 we define the baseline hazard. The baseline hazard is the hazard function for a reference individual who has all covariates set to their referenc e value .

Better explanation of the theory

- $h(t)$: The hazard rate for an individual with covariates x at time t. Is the instantaneous risk of having the event at time t, given that the person has survived up to t.
- $h_0(t)$: the baseline hazard is the hazard when all covariates are their reference value (0). It plays the role of an “intercept function” and describes how the risk changed over time for the reference individual.

- $x_1, ..., x_k$: are the explanatory variables they can be continous or categorical, and they may be fixed or time dependent.
- $\beta_1, ..., \beta_k$ are the regression koefficients β_j describes the effect of the covariate x_j on the hazard. The quantify $\exp(\beta_j)$ is a hazard ratio: it tells us how many times larger or smallaer the hazard become if x_j increases by one unit, holding other covariates constant.

How are the coefficients estimated?

The coefficients are estimated by **partial maximum likelihood,** which uses the order and risk sets of event times and does not require specifying the form of $h_0(t)$. The exacty shape of $h_0(t)$ is usually not of primary interest; we mainly care about the hazard ratios $\exp(\beta_j)$

Proportional Hazard property

The Cox model is a proportional hazards model; this means that the covariates are multiplicatively related to the hazard and the hazard ratio is constant over time.

Relative survival

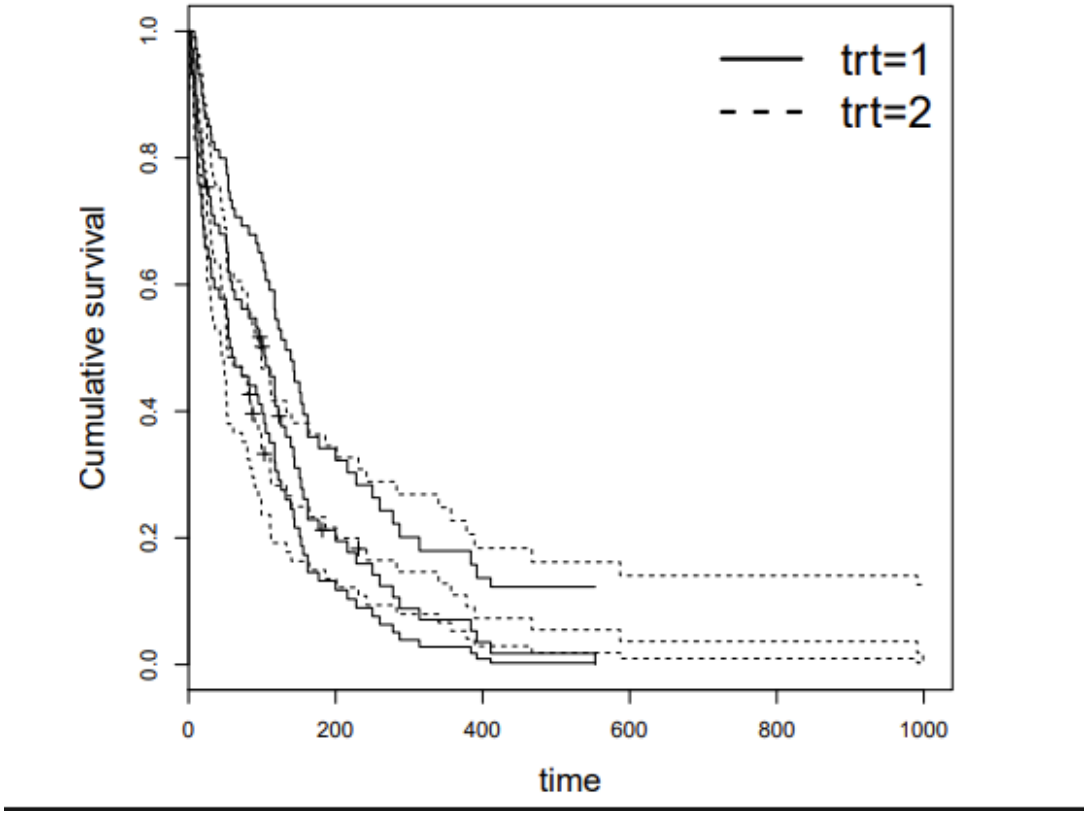
In relative survival, the total hazard can be written has:

$$h(t) = h_{background}(t) + h_{excess}(t)$$

Where

- $h_{background}(t)$is the expected hazard in the general population (same age, sex, etc).
- $h_{excess}(t)$ is the extra hazard due to the disease or condition of interest (often modelled with Cox Regression).

Cox Regression (KM)



This plot showns the KM curve for veteran lung cancer trial by split by treatment group. On the x-axis is time and y-axis presents the cumulative survival: this is the estimated survival function. Which is the probablity that a patient is still alive after time t.

What does the curves show?

The curves show:

- $trt = 1$: the patients who got the standard treatment
- $trt = 2$: the patients who got the test (experimental) treatment.

Each line is the Kaplan-Meier estimate of survival in that group, with thin lines around it showing 95% confidence intervals.

So for each time t

- The height of the solid curve = estimates probablity of a standard treatment patient is sitll alive after t.

- The height of the dashed curve = same probablity for a test-treatment patient.

How do we interpret this?

Visually we can say that:

- Both curves drop quickly → poor overall prognosis (high early mortality).
- At many times, one curve is above the other:
  - If the dashed line is $(trt = 2)$ above the solid line then the treatment seems to have better survival at those times (higher probablity to still be alive).
  - If they cross or are close the dfference is smaller or unclear

The confidence bands show the uncertainty around each KM curve. Where the bands are wide, we have fewer patients left and less precise estimates.

Output of the Cox Regression

```
Call:
coxph(formula = v.times ~ karno + trt + celltype, data = veteran)

  n= 137, number of events= 128

                      coef exp(coef)  se(coef)      z Pr(>|z|)
karno             -0.031271  0.969213  0.005165 -6.054 1.41e-09 ***
trt2               0.261744  1.299194  0.200923  1.303  0.19267
celltypesmallcell  0.824980  2.281836  0.268911  3.068  0.00216 **
celltypeadeno      1.153994  3.170833  0.295038  3.911 9.18e-05 ***
celltypelarge      0.394625  1.483828  0.282243  1.398  0.16206
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

                  exp(coef) exp(-coef) lower .95 upper .95
karno                0.9692     1.0318    0.9595    0.9791
trt2                 1.2992     0.7697    0.8763    1.9262
celltypesmallcell    2.2818     0.4382    1.3471    3.8653
celltypeadeno        3.1708     0.3154    1.7784    5.6534
celltypelarge        1.4838     0.6739    0.8534    2.5801

Concordance= 0.737  (se = 0.03 )
Rsquare= 0.36   (max possible= 0.999 )
Likelihood ratio test= 61.07  on 5 df,   p=7.307e-12
Wald test            = 63.41  on 5 df,   p=2.397e-12
Score (logrank) test = 66.55  on 5 df,   p=5.347e-13
```

In the ouput above we can see the cox regression being performed. The predictors that are included are:

- karno = Karnofsky performance score (higher = better health).
- trt2 = treatment indicator (1 = special treatment, 0 = standard).
- celltype: categorical with 4 levels (smallcell, adeno, large, squamous is the reference).

Number of patients are 137 and number of deaths (events) are 128. This means that alomst all of the people studied died during follow up.

Interpretation of the coefficients

- **Hazard ratio for Karno = 0.97** per 1 point unit increase for the Karnofsky score (remember that we take the exp() of the coef).
    - We can say that for each additional point of health score, hazard (risk of death) decrases by 3% which is very significant ($p<0.0001$).
- **Hazard ratio for trt2 = 1.30** which means that group 2 has 30% higher hazard than roup 1. But this is not statistically significant meaning we cannot say there is evidence that treatment 2 is better or worse. **This matches what the KM plots often show.**
- **Hazard Ratio of smallcell = 2.28** which means that there is a 2.28 times the hazard of the baseline group (sqamoud). This is an aggressive cancer which makes clinically sense.
- **Hazard ratio of aden = 3.17** which means that patient with adenocarcinoma have 3.17 times the hazard relative to squamous.
- **Hazard ratio large = 1.48** which means that large cell cancet patient have 1.48 the hazard, but not significant.
- **Concordance = 0.737.** This is like the AUC for survival data. 0.73 is fairly good predictive discrimination.
- **Likelihood ratio, Wald, Score (logrank) test**
    - All have p-values that are low. meaning that the model is highly significant. This means that the covariates as a group strongly predict survival

Should we remove a variable because it’s not significant?

In Cox regression (and regression in general); we should not simply remove a variable just because its p-value is > 0.05 (for example, the treatment indicator trt2). A non-signficiant reslut does njot mean “no effects”; it can be due to limited sample size, high variability, or confounding. Dropping important variables — especially key design variables like treatment can bias the estimated effects of the other covariates and lead to wrong scientific conclusions.

Instead we should keep scientifically important variables (such as treatment, age, sex, known risk factors) in the model regardless of their p-values. We can then:

- Check the model assumption (e.g proportional hazards)
- Consider interactions.
- Compare full and reduced models with likelihood ratio test rather than doing automatic p-value base variable deletion.

Model fitting and comparing models

R2

- Concordance is similar to the idea of R2 in linear regression but not the same.
- R2 measures how well the model explains the variation in survival.
- High R2 → better fit.
- Lower R2 → worse fit.
- However in survival models R2 is not very reliable so we rarely use it alone.

Concordance

Concordance is the most important measure of model fit in survival analysis.

- It measures how well the model orders individuals by risk.
    - Values:
        - 0.5 → no better than random guessing.
        - 0.7 → acceptable discrimination.
        - 0.8+ → strong model.
- Higher concordance means better model fit.

Interpretation

> If the model predicts patient A has higher risk than patient B, how often is that correct?

Then you can say “the lower the concordance, the worse the model”.

Information Criterions (AIC/BIC)

When comparing Cox regression models we can simply look at validation criterions such as AIC or BIC. The AIC is formulated as follows:

$$AIC = -2\ell + 2K \quad AIC_c = AIC + \frac{K(K+1)}{n-K-1}$$
$$BIC = -2\ell + K\log n$$

Where

- L = log likelihood
- K = number of parameters in the model
- n = sample size.

Delta AIC (ΔAIC)

Each model has an AIC value — the lower AIC equals the better model fit. But RAW AIC values themselves dont mean anything. We need a good way to compare models. Let’s say that we have several models:

- Model A: AIC = 200
- Model B: AIC = 202
- Model C: AIC = 210.

First find the best model (the one with smallest AIC):

$$AIC_{\min} = 200$$

The calculate the ΔAIC for each model:

$$\Delta AIC_i = AIC_i - AIC_{\min}$$

- **Model A:** $\Delta AIC = 200 - 200 = 0$
- **Model B:** $\Delta AIC = 202 - 200 = 2$
- **Model C:** $\Delta AIC = 210 - 200 = 10$

How to interpret delta AIC

So ΔAIC tells you much worse each model is compared to the best one. You can think of intepreting the resluts as:

- $\Delta AIC = 0$ → best supported model
- $\Delta AIC \leq 2$ → almost as good as the best
- $\Delta AIC 4 - 7$ → mcuh less support / weaker
- $\Delta AIC = 10$ → essentially no support → model is bad.

Model validation of Cox regression

In a Cox regression model the proportional hazard assumption says:

> The effect of each covariate on the hazard (i.e the β's) is constant over time.

In other words the hazard ratio for treatment, age, etic, should not change with time. To check this we can use **Scheonfeld** residuals where:

> If the covariate effect changes with time → proportional hazards is violated.

Schoenfeld residuals

For each event time (for each person who fails) and for each covariate c, we compute a residual:

$$r_{ci} = x_{ci} - \sum_{k \in R_i} x_{ck} p_{ck}$$

Where

- i : the individual who fails at that event time.
- c : a particiular covariate (e.g treament, age, etc).
- x_{ci} : the actual value of covariate c for the failing individual

    - Example if c is treament (0/1) and a person i got treatment then $x_{ci} = 1$
- R_i : the risk set at the time individual i fails = all individuals who are still at risk just before failiure time.
- For each person k in the risk set R_i:

$$p_k = \frac{\exp(\beta^T X_k)}{\sum_{e\ in R_i} \exp(\beta^T X_r)}$$

    - this is the model based probablity that person k would fail at that time (according to the fitted Cox model).

  Then:

$$\sum_{k \in R_i} x_{ck} p_k$$

is the expected value of covariate c at that time, given that soemone from the risk set fails, under the Cox model.

So the Scoenfeld residual is:

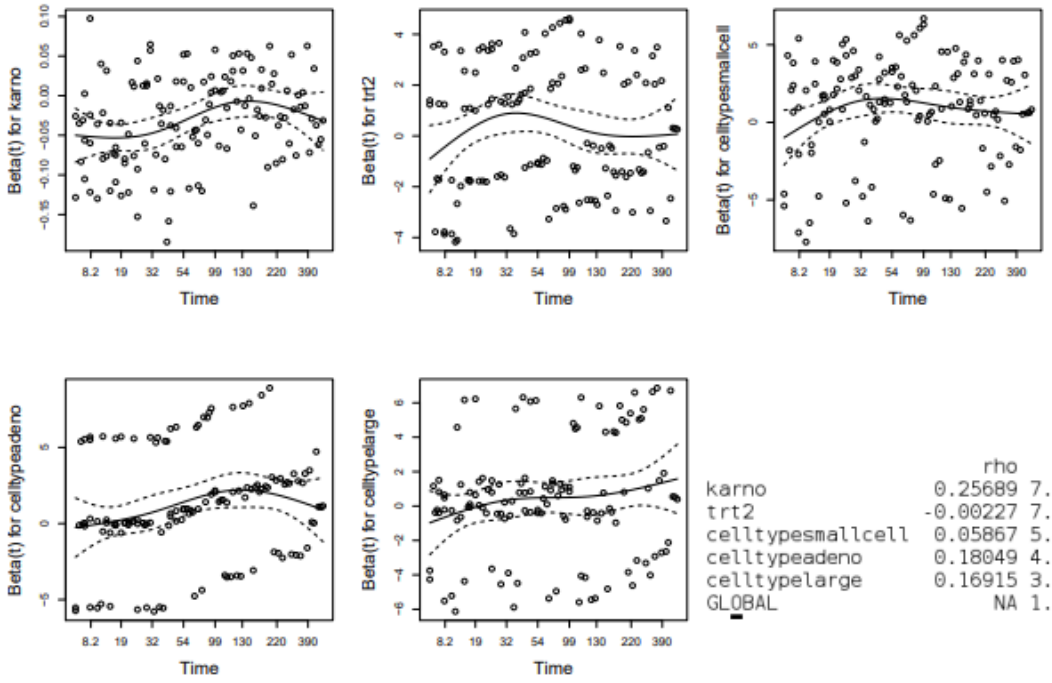
> Observed covariate value for the failing individual minus expected covariate value (averaged over the risk set, weighted by model probablities)
> d $r_{ci} = \text{observed } x_{ci} - \text{expected } x_c \text{ at that even time}$

IMPORTANT NOTATION

The Schoenfeld residuals are **only defined** for unscensored individuals i.e those that failed. And they should be independent of time. In other words the slope of the regression of the residuals on time should be 0.

Schoenfeld plot and output

▸ Tests of Proportionality in SAS, STATA and SPLUS



	rho	chisq	p
karno	0.25689	7.92e+00	0.00489
trt2	-0.00227	7.91e-04	0.97756
celltypesmallcell	0.05867	5.16e-01	0.47250
celltypeadeno	0.18049	4.73e+00	0.02968
celltypelarge	0.16915	3.92e+00	0.04778
GLOBAL	NA	1.76e+01	0.00353

Each panel check whether one covariate violates the proportional hazard assumption (PH):

- Circles = Raw Schoenfeld residuals.
- Solid curve = smooth trend over time.
- Dotted curved = confidence band.
- A horizontal line (slope =0) = PH assumption OK.
- Clear trend (uppward/downward or curved) = PH violated.

The following plots for the covariates can be read as:

- **Karno:** The smoothed line increases over time. It is not flat → meaning that karno changed with time. P value is as well <0.05 which we can interpret as **patients** performance score influences hazard differently at early vs lates times → PH assumption is violated.
- **TR2:** The smoothed curve is almost flat with no visible upward or downard trend. The p-value i > 0.05 meaning that the assumption is not violated. The

effect of treatment is constant over time.

- **Small cell** : no violation
- **Adeno:** violated
- **Large :** no violation
- **GLOBAL TEST** : the global test combine all covariates together. The p-value is less than 5 meaning that at least one covariate is violating the PH assumption.
    - Karno is probably doing that.

DESICON OF VIOLATION

We want p-values > 0.05 and rougly flat smooth curves → PH assumption reasonable. If p-values are < 0.05 + visible trend → investigate and possibly adjust the model.

Extending the proprtional hazards model

The basic assumption is that the hazard ratio between different groups is constant with time.

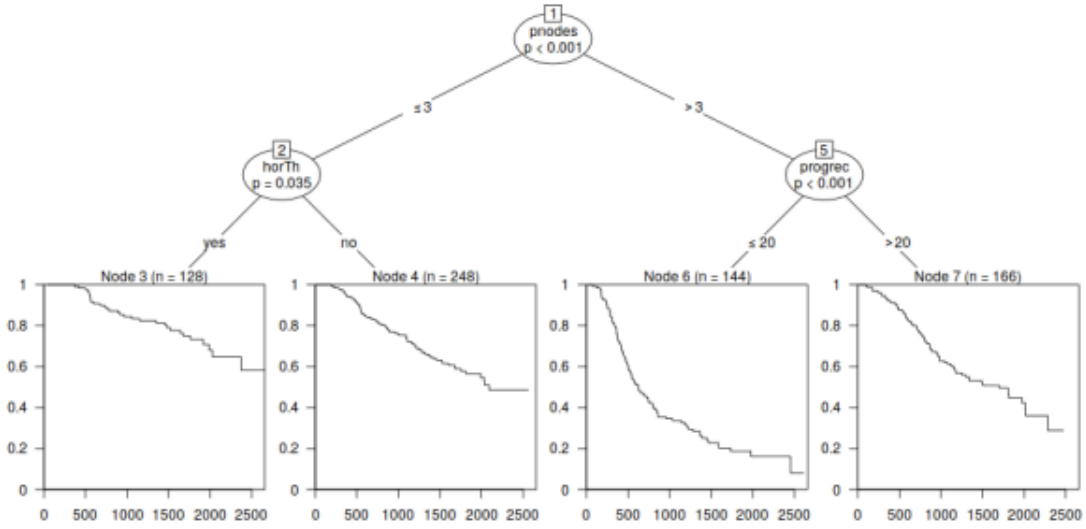
The Cox model assumes proportional hazards, i.e that the hazard ratio between groups are constant over time. When this is violated, we can allow the effect of a covariate to depnd on time. In the discrete case (covariate changes at known times), each follow-up is split into intervals with a constant value. In the continuous case, we specify a parametric time interation, e.g:

$$h(t) = h_0(t)\exp(\beta x_1 + ... + \beta_k x_k + c x_{k+1} t)$$

So that the effect of x_{k+1} varies with time.

Decision trees

GBSG2: German Breast Cancer Study Group 2
pnodes: number of positive nodes;
horTh: hormonal therapy; progrec: progesterone receptor
time: recurrence free survival time in days



This slide shows a survival desicion tree (also called a survival tree or recursive partitioning tree) fitted to the GBSG2 breast cancer data. Lets explain clearly what is happened and how to interpret the tree and the survival curves.

What a Survival Tree Does

A survival tree:

- Splits the population into subgroups (nodes).
- Used covariates to from groups that differ in survival.
- At each split, chooses the covariate and cutoff that best separates survival curves (using log-rank test or similar).
- Produces a tree where each terminal node has its own Kaplan Meier curve.

This is the survival version of CART decision tree.

What the variables mean

- pnodes = number of positive lymph nodes → a strong predictor of cancer recurrance.
- horTH = hormonal therapy)yes/no)
- progrec = progesterone receptor level
- time = recurrence free-survival time (days)

Interpretation of the Tree

First split $pnodes \leq 3$ has a strong significant o-value. This is the strongest predictor:

- Patient with ≤ 3 positive nodes go to the lft.
- Patient with > 3 positive nodes go to the right

This basically means that having many cancer positive lymh nodes is asosciated with much worse survival.

The next split uses **horTh**

- **Yes** (hormonal therapy) → node 3
- No (No hormonal therapy) → node 4

This can be interpreted as among patient with few positive nodes, hormonal therapy improves survival. You can see that node 3 has higher KM curve.

The right branch **progesterone receptor level (progrec)** with a cutoff 20 says

- progrec ≤ 20 → node 6
- progrec > 20 →node 7

This can be interpreted as among patient with positive nodes, a higher progesterone recepter level is associated with better prognosis, where node 7 has noticeably better survival than node 6.

How to read the Kaplan Meier curves

- **Node 3:**

    - Best prognosis in the entire tree
    - Very high survival curve
- **Node 4**
    - Worse than node 3, buit still decent survival because pnodes are low.
- **Node 6**
    - Worst survival group
    - Steep drop → high recurrence risk.
- **Node 7**
    - Better than node 6, but still worse than the ≤ 3 node groups

What does the tree tell us?

The main points are:

1. Number of positive nodes (pnodes) are the strongest predictor of recurrence (split at pnodes = 3).
2. Among patients with few nodes, hormonal therapy improves prognosis.
3. Among patients with many nodes, high progesterone receptr levels improve prognosis.

However, clinically, lymph node involvement is the main risk factor. Thearpy and hormone receptor status refine the risk groups.

So why use survival trees?

- Handles interactions automatically
- Non-parametric (no PH assumption like Cox regression)
- Provides easto-to-explain subgroups for clinicans.
- Helps identify risk strata for personalized treatment.