Survival analysis : not so far away from linear regression. But here we are interested in how risk factors affects time to disease (and not how assiocated with presence or absence of disease).

Here the outcome of interest is the TIME TO AN EVENT. The event is adverse (death). In survival analysis the response is then a time (continuous but measured discretely).

One special characteristic of survival data is CENSORING : incomplete response (we know their survival time t but don't know the exact time).

If there was no censoring we could use standard regression procedure (predict time before death).

BUT not too interesting because we are more interested in the probability of surviving past a given time than just the expected time of event.

In survival analysis we use the HAZARD FUNCTION : $h(t) = h_0(t) \exp(\beta x)$

Censoring : when the waiting time exceeds the observation time. For the survival analysis methods to be valid, the censoring mechanism must be independent of the survival mechanism.

We observed right :

- Fixed type I : sample of n units is followed for a fixed time tau : the number of units experiencing the event is random but the total duration of the study is fixed.
- Fixed type II : a sample of n units is followed as long as necessary until d units have experienced the events.
- Random censoring : unit associated with a potential censoring time Ci and a potential life time. &I indicates if the observation I is terminated by death or by censoring.

In all these schemes, the censoring mechanism is non-informative and they all lead to essentially the same likelihood function.

We observed two possibilities : & = 0 the event was censored and & = 1 the event was observed.

Survival function : The survival function is given by : S(t) = Pr(T >t) = 1 – F(t) where F(t) is the cumulative distribution function for T. It gives the probability that a subject will survive past time t.

Characteristics : non-increasing, S(0) = 1

In theory the survival function is smooth but in practice we observe events on a discrete time scale.

When we want to estimate S(t), we assume that every subject follows the same survival function. In case of no censoring we just have to use a simple-non parametric estimator (number of individuals that survive until t)/total number of subject).

BUT in case of censoring we cannot use this procedure, then we use the KAPLAN-MEIER estimator. Kaplan-Meier estimator is also called product limit estimator. It is graphically shown with CONFIDENCE BANDS and has the form of a down staircase.

When there is no censoring, the Kaplan-Meier curve is equivalent to the empirical distribution.

It is possible to test for differences between curves using the LOG-RANK test.

The null hypothesis is : equality of survival curves. We compute the expected number of deaths for each unique death time in the data, assuming that the chance of dying for subjects at risk is the same for each group. Then we do the total number of expected deaths for each group as the sum of the expected numbers for each time !
Then the test compared the observed number of deaths in each group to the expected number using a chi-squarred test.

In survival analysis we often want to assess which time periods have high or low chances of failure among those still at risk at the time. We can characterize these risk using the instantaneous failure rate or hazard function h(t) : It is the probability that an individual experiences the event in a small time interval s, given that the individual has survived up to the beginning of the interval, when the size of the time interval approaches 0.

The hazard function and survivor function can be related together : $S(t) = \exp(-H(t))$.

Where H(t) is the hazard function integrated on the interval. We can estimate the hazard function as the proportion of individuals experiencing the event in an interval per unit time, given that they have survived to the beginning of the interval.

These estimators are NON-PARAMETRIC : we do not make any assumptions about the distributional form of the survival time T. Some common parametric models for survival data is : exponential, gamma, log-normal. If the parametric model is correct we can then estimate S(t) more precisely.

Cox regression : regression technique to deal with censored survival data. It model the hazard function. It is carried out similarly to regression but with linearity assumed on the log hazard scale.