Date :

UIT2502 - Data Analytics & Visualization

Assignment - 2

BY: R. NITHYASRI
REG NO: 3122225002086
CLASS : IT - 'B'.

1) What is the difference between censored and uncensored data in survival analysis? Provide examples of each.

1) Censored data:- Censored data occurs when we do not know the exact time of the event for some subjects, but we do know that they survived or persisted up to a certain point. This usually happens for one of the following reasons:

Right censoring :- It happens when the subject hasn't experienced the event by the end of the study period or has dropped out before the study ends.

Left censoring :- When we know the event happened before a certain time but don't know exactly when.

Interval censoring :- When the event happens within a known time interval but not at an exact time.

Date :

Eg:- In a clinical trial, if a patient does not relapse before the study ends, their time is recorded up to the study's end date, the data is right-censored.

If a participant drops out of a study for personal reasons, their data would also be right-censored since we only know they survived the dropout date.

2) Uncensored data:- We know the exact time at which the event occured for a subject.

Eg:- In a study tracking time to recovery after a surgery, if a patient fully recovers within the study period and the exact recovery time is recorded.

In a job satisfaction study tracking employee turnover, if an employee leaves the company during the observation period, their exact tenure duration is uncensored data.

Date :

2) Survival Functions :

Explain the Kaplan-Meier estimator and its use in estimating the survival function. How does it handle censored data?

The Kaplan-Meier estimator is a non-parametric tool used in survival analysis to estimate the survival function $S(t)$, which is the probability that a subject survives beyond a given time t. It is particularly useful for handling right-censored data.

Key points :-

1) Calculation :- The estimator calculates survival probabilities at each event time t_i, using only those subjects who are still 'at risk'.

2) Censored Data Handling :- Censored subjects contribute to the risks set until the time they are censored.

3) Kaplan-Meier Curve :- The results are visualized as a stepwise curve that decreases at each event time, with small vertical ticks marking censored data points.

Date :

Kaplan - Meier Survival Function :-

$$S(t) = \prod_{t_i \leq t} \left(1 - \frac{d_i}{n_i}\right)$$

t_i = Time of the i^{th} event

d_i = Number of events at t_i

n_i = Number of subjects at risk before time t_i.

3) ~~Hazd~~ Hazard functions :-

The hazard function, $h(t)$ represents an instantaneous rate at which events occur at a particular time t, given that the individual has survived up to that time.

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

Relationship to survival function,

$$S(t) = e^{-\int_0^t h(u)\,du}$$

A higher hazard implies higher risk of the event occuring.

Date :

A constant hazard suggests a steady risk over time.

A decreasing hazard over time might occur in scenarios where early failures are more common but those who survive initially face reduced risk later.

4) Proportional hazards :-

The proportional hazards assumption in the Cox Proportional Hazards model states that the hazard ratio between individuals with different covariates is constant over time. This means that the effect of each predictor on the hazard remains the same throughout the study period.

Testing for the Assumption :-

1) Schoenfeld Residuals

2) Log - Log Survival Curves

3) Time - Dependent Covariates.

Date :

5) Time-varying covariates :-

Time-varying covariates in survival analysis can be allowing covariates to change over time, typically through an extended Cox model or splitting time into intervals. Challenges include increased model complexity, difficulty in interpreting hazard ratios, and potential violations of the proportional hazards assumptions.

Relation between hazard function and covariates :-

$$h_i(t) = h_0(t)\, e^{\beta_1 X_{i1}(t) + \beta_2 X_{i2}(t) \dots}$$

where $X_{ij}(t)$ represents the value of covariate j at time t for individual i.