Survival Analysis: Time To Event Modelling

Sudipta Das

Assistant Professor,
Department of Computer Science,
Ramakrishna Mission Vivekananda Educational & Research Institute

Outline I

1. Introduction
   - Areas of application
   - Special features of Survival data
   - Types of modeling in Survival Analysis

Introduction

Survival Analysis, the problem of analyzing time to event data arises in a number of applied fields, such as

- Medicine,
- Biology,
- Public health,
- Epidemiology,
- Engineering,
- Economics,
- Demography,
- Banking.

- Remission Duration from a Clinical Trial for Acute Leukemia
  - Freireich et al. (1963) report the results of a clinical trial of a drug 6-mercaptopurine (6-MP) versus a placebo in 42 children with acute leukemia.
  - The trial was conducted at 11 American hospitals.
  - Patients were selected who had a complete or partial remission of their leukemia induced by treatment with the drug prednisone.
    - A complete or partial remission means that either most or all signs of disease had disappeared from the bone marrow.
  - The trial was conducted by matching pairs of patients at a given hospital by remission status (complete or partial) and randomizing within the pair to either a 6-MP or placebo maintenance therapy.
  - Patients were followed until their leukemia returned (relapse) or until the end of the study (in months).

Introduction: Medicine II

Table: Remission duration of 6-MP versus placebo in children with acute leukemia

Pair	Remission Status at Randomization	Time to Relapse for Placebo Patients	Time to Relapse for 6-MP Patients
1	Partial Remission	1	10
2	Complete Remission	22	7
3	Complete Remission	3	32+
4	Complete Remission	12	23
5	Complete Remission	8	22
6	Partial Remission	17	6
7	Complete Remission	2	16
8	Complete Remission	11	34+
9	Complete Remission	8	32+
10	Complete Remission	12	25+
11	Complete Remission	2	11+
12	Partial Remission	5	20+
13	Complete Remission	4	19+
14	Complete Remission	15	6
15	Complete Remission	8	17+
16	Partial Remission	23	35+
17	Partial Remission	5	6
18	Complete Remission	11	13
19	Complete Remission	4	9+
20	Complete Remission	1	6+
21	Complete Remission	8	10+

Introduction: Medicine III

- Times to Death for a Breast-Cancer Trial
  - In a study (Sedmak et al., 1989) designed to determine if female breast cancer patients, originally classified as lymph node negative by standard light microscopy (SLM), could be more accurately classified by immunohistochemical (IH) examination of their lymph nodes with an anticytokeratin monoclonal antibody cocktail, identical sections of lymph nodes were sequentially examined by SLM and IH.
  - The significance of this study is that 16% of patients with negative axillary lymph nodes, by standard pathological examination, develop recurrent disease within 10 years.
  - Forty-five female breast-cancer patients with negative axillary lymph nodes and a minimum 10-year follow-up were selected from The Ohio State University Hospitals Cancer Registry.
  - Of the 45 patients, 9 were immunoperoxidase positive, and the remaining 36 remained negative. Survival times (in months) for both groups of patients are given.

Introduction: Medicine IV

Table: Times to death (in months) for breast cancer patients with different immunohistochemical responses

Immunoperoxidase Negative	19, 25, 30, 34, 37, 46, 47, 51, 56, 57, 61, 66, 67, 74, 78, 86, 122+, 123+, 130+ , 130+, 133+, 134+, 136+, 141+, 143+, 148+, 151+, 152+, 153+, 154+, 156+, 162+, 164+, 165+, 182+,189+
Immunoperoxidase Positive	22, 23, 38, 42, 73, 77, 89, 115, 144+

Introduction: Medicine V

- Death Times of Psychiatric Patients
  - Woolson (1981) has reported survival data on 26 psychiatric inpatients admitted to the University of Iowa hospitals during the years 1935–1948.
  - This sample is part of a larger study of psychiatric inpatients discussed by Tsuang and Woolson (1977).
  - Data for each patient consists of age at first admission to the hospital, sex, number of years of follow-up (years from admission to death or censoring) and patient status at the follow-up time.

Introduction: Medicine VI

Table: Survival data for psychiatric inpatients

Gender	Age at admission	Time of Follow-up		Gender	Age at admission	Time of Follow-up
Female	51	1		Female	30	37+
Female	58	1		Female	33	35+
Female	55	2		Male	36	25
Female	28	22		Male	30	31+
Male	21	30+		Male	41	22
Male	19	28		Female	43	26
Female	25	32		Female	45	24
Female	48	11		Female	35	35+
Female	47	14		Male	29	34+
Female	25	36+		Male	35	30+
Female	31	31+		Male	32	35
Male	24	33+		Female	36	40
Male	25	33+		Male	32	39+

Introduction: Banking I

- To a bank, probably the most important thing is the loans it provides.
  - Loans are the way through which a bank makes money.
- Consequently, when loans go bad, it can be fatal to a bank.
  - Resulting in Bank failure.
  - Sometimes, the government is forced to step in and bail out the bank, costing taxpayers' money.
- Provide loans judicially.
  - Based on customer profile, carry an quantitative assessment of the risk of non-repayment

Introduction: Banking II

- Example 1:- German banks credit (part) data

Duration	Amount	Installment Rate in %	Age	No. of Credits	No. of people Maintenance	$\cdots$ $\cdots$	Type
6	1169	4	67	2	1	$\vdots$	Good
48	5951	2	22	1	1	$\vdots$	Bad
12	2096	2	49	1	2	$\vdots$	Good
42	7882	2	45	1	2	$\vdots$	Good
24	4870	3	53	2	2	$\vdots$	Bad
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
45	1845	4	23	1	1	$\vdots$	Bad
45	4576	3	27	1	1	$\vdots$	Good

Introduction: Banking III

- Data contains two kinds of loan information: Good and Bad
- A binary question: Can I say, whether the loan is good or bad, considering the customer profile
- Answer: Yes, by standard classification techniques (e.g. logistic regression, random forest, etc.)
- Question: When would the loan go bad?
- Question: Between two banks which one is possessing riskier loans?
  - No answer by classification method.
- Answer is given by Survival Analysis

Introduction: Different names in different sectors I

- Banking/Economics: *Time to Event Modeling*
- Engineering: *Reliability Analysis*
- Biological/Medical Science: *Survival Analysis*

Introduction: Different names in different sectors II

- We model the time to failure [T] data (or time-to-event data or survival data)
  - Time **from** *medication* **to** *relapse of leukemia*
    - When the leukemia will returned for a patient?
  - Time **from** *date of breast cancer detection* **to** *death of patient*
    - When the patient will die?
  - Time **from** *date of admission* **to** *death of patient*
    - When the patient will die?
  - Time **from** *loan approved* **to** *loan default*
    - When the customer will be defaulter?

Two special features of Survival data I

- Non-negativeness
  - Simple regression fails
- Incompleteness
  - Data set contains either *censored* or *truncated* observations

Two special features of Survival data II

- Censored data arises when an individual's life length is known to occur only in a certain period of time.
  - *Right censoring*: where all that is known is that the individual is still alive at a given time,
  - *Left censoring*: when all that is known is that the individual has experienced the event of interest prior to the start of the study, or
  - *Interval censoring*: where the only information is that the event occurs within some interval.

Two special features of Survival data III

- Truncation schemes are
  - *Left truncation*: where only individuals who survive a sufficient time are included in the sample and
  - *Right truncation*: where only individuals who have experienced the event by a specified time are included in the sample.

Two types of modeling in Survival Analysis I

- The duration of failure [T] is modeled
  - Survival function Modeling
  - Hazard/failure rate Modeling
- Note: Two model types are related

Two types of modeling in Survival Analysis II

- Survival function Model:
  - Probability of no failure till time t,

$$
\begin{aligned}
P(T > t) &= \int_t^\infty f(u)du \\
&= S(t).
\end{aligned}
$$

Two types of modeling in Survival Analysis III

- Hazard/failure rate Model:
  - Instantaneous rate of failure given survival till t,

$$
\begin{aligned}
h(t) &= \lim_{\Delta t \to 0} \frac{P(t \leq T < t + \Delta t | T > t)}{\Delta t} \\
&= \frac{f(t)}{P(T > t)} \\
&= \frac{f(t)}{\int_t^\infty f(u)du}
\end{aligned}
$$

  - Cumulative Hazard:

$$H(t) = \int_0^t h(u)du$$

Two types of modeling in Survival Analysis IV

- Relation between *Survival function* and *Hazard Function*

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