A Complete Guide To Survival Analysis In Python, part 1

[<= Previous post]
[Next post =>]

Tags: Python, Statistics, Survival Analysis

This three-part series covers a review with step-by-step explanations and code for how to perform statistical survival analysis used to investigate the time some event takes to occur, such as patient survival during the COVID-19 pandemic, the time to failure of engineering products, or even the time to closing a sale after an initial customer contact.

---

comments

By [Pratik Shukla], Aspiring machine learning engineer.

Survival Analysis Basics

Survival analysis is a set of statistical approaches used to find out the time it takes for an event of interest to occur. Survival analysis is used to study the **time** until some **event** of interest (often referred to as **death**) occurs. Time could be measured in years, months, weeks, days, etc. The event of interest could be anything of interest. It could be an actual death, a birth, a retirement, etc.

How it can be useful to analyze ongoing COVID-19 pandemic data?

(1) We can find the number of days until patients showed COVID-19 symptoms.

(2) We can find for which age group it's deadlier.

(3) We can find which treatment has the highest survival probability.

(4) We can find whether a person's sex has a significant effect on their survival time?

(5) We can also find the median number of days of survival for patients.

We are going to perform a thorough analysis of patients with lung cancer. Don't worry once you understand the logic behind it, you'll be able to perform it on any data set. **Exciting, isn't it?**

Survival analysis is used in a variety of field such as:

- Cancer studies for patients survival time analyses.
- Sociology for "event-history analysis".
- In Engineering for "failure-time analysis".
- Time until product failure.
- Time until a warranty claim.
- Time until a process reaches a critical level.
- Time from initial sales contact to a sale.
- Time from employee hire to either termination or quit.
- Time from a salesperson hire to their first sale.

In **cancer studies**, typical research questions include:

(1) What is the impact of certain clinical characteristics on patient's survival? For example, is there any difference between the group of people who has higher blood sugar and those who don't?

(2) What is the probability that an individual survives a specific period (years, months, days)? For example, given a set of cancer patients, we will be able to tell that if 300(random number) days after the diagnosis of cancer has been passed, then the probability of that person being alive at that time will be 0.7 (random number).

(3) Are there differences in survival between groups of patients? For example, let's say there are 2 groups of people diagnosed with cancer. Those 2 groups were given 2 different kinds of treatments. Now our goal here will be to find out if there is a significant difference between the survival time for those 2 different groups based on the treatment they were given.

Objectives

In cancer studies, most of the survival analyses use the following methods.

(1) *Kaplan-Meier plots* to visualize survival curves.

(2) *Nelson-Aalen plots* to visualize the cumulative hazard.

(3) *Log-rank test* to compare the survival curves of two or more groups

(4) *Cox proportional hazards regression* to find out the effect of different variables like age, sex, weight on survival.

Fundamental concepts

Here, we start by defining fundamental terms of survival analysis, including:

- Survival time and event.
- Censoring of data.
- Survival function and hazard function.

Survival time and type of events in cancer studies

Survival Time: referred to an amount of time until when a subject is alive or actively participates in a survey.

There are mainly three types of events, including:

(1) **Relapse**: a deterioration in someone's state of health after a temporary improvement.

(2) **Progression**: the process of developing or moving gradually towards a more advanced state. (Improvement in health.)

(3) **Death**: the destruction or permanent end of something.

Censoring

As mentioned above, survival analysis focuses on the occurrence of an event of interest (e.g., birth, death, retirement). But there is still a possibility that the event may not be observed for various reasons. Such observations are known as censored observations.

Censoring may arise in the following ways:

1. A patient has not (yet) experienced the event of interest (death or relapse in our case) within the study period.
2. A patient is not followed anymore.
3. If a patient moves to another city, then follow-up might not be possible for the hospital staff.

This type of censoring, named *right censoring*, is handled in survival analysis.

There are three general types of censoring, right-censoring, left-censoring, and interval-censoring.

Right Censoring: The death of the person.

Left Censoring: The event can't be observed for some reason. It includes events that occurred before the experiment started. (e.g., number of days from birth when the kid started walking.)

Interval Censoring: When we have data for some intervals only.

Survival and hazard functions

We generally use two related probabilities to analyse survival data.

(1) The survival probability

(2) The hazard probability

To find survival probability, we'll be using survivor function S(t), which is the Kaplan-Meier Estimator. Survival probability is the probability that an individual (e.g., patient) survives from the time origin (e.g., diagnosis of cancer) to a specified future time t. For example, S(200) = 0.7 means that after 200 days has

passed since the diagnosis of cancer, the patient's survival probability has dropped to 0.7. If the person stays alive at the end of an experiment, then that data will be censored.

The hazard probability, denoted by h(t), is the probability that an individual(e.g., patient) who is under observation at a time t has an event(e.g., death) at that time. For example, If h(200) = 0.7, then it means that the probability of that person being dead at time t=200 days is 0.7.

Note that, in contrast to the survivor function, which focuses on not having an event, the hazard function focuses on the event occurring. I think we can clearly see that higher survival probability and lower hazard probability is good for the patient.

Let's move forward to the cool coding part!

You can download the dataset from **here**.

Data Description

inst : Institution code

time : Survival time in days

status : Censoring status ; 1 = censored ; 2 = dead

age : Age in years

sex : Male = 1 ; Female = 2

ph.ecog : Ecog performance score ; 0= good ; 5 = dead

ph.karno : Karnofsky performance score (bad=0 ; good=100) rated by physician

pat.karno : Karnofsky performance score as rated by patient

meal.cal : Calories consumed at meals

wt.loss : Weight loss in last six months

Kaplan-Meier Estimator

The **Kaplan–Meier estimator** is a non-parametric statistic used to estimate the survival function (probability of a person surviving) from lifetime data. In medical research, it is often used to measure the fraction of patients living for a certain amount of time after treatment. For example, Calculating the amount of time(year, month, day) certain patient lived after he/she was diagnosed with cancer or his treatment starts. The estimator is named after **Edward L. Kaplan** and **Paul Meier**, whom each submitted similar manuscripts to the *Journal of the American Statistical Association*.

The formula for Kaplan-Meier is as follows:

The probability at time ti, S(ti), is calculated as

$$\widehat{S(t)} = \prod_{t_i<t} \frac{n_i - d_i}{n_i}$$

We can also write it as

$$S(t_i) = S(t_{i-1}) * \left(1 - \frac{d_i}{n_i}\right)$$

Where,

$S(t_{i-1})$ = the probability of being alive at t_{i-1}

n_i = the number of patients alive just before t_i

d_i = the number of events at t_i

t_0 = 0

$S(0)$ = 1

Survival Function

For example,

$$S(1) = S(0) * \left(1 - \frac{d_1}{n_1}\right) = \left(1 - \frac{d_1}{n_1}\right)$$

$$S(2) = S(1) * \left(1 - \frac{d_2}{n_2}\right)$$

$$S(3) = S(2) * \left(1 - \frac{d_3}{n_3}\right)$$

In a more generalized way, we can say that,

$$S_t = \frac{Number\ of\ subjects\ at\ risk\ at\ the\ start\ -\ Number\ of\ subjects\ that\ died}{Number\ of\ subjects\ at\ risk\ at\ the\ start}$$

Survival function simplified.

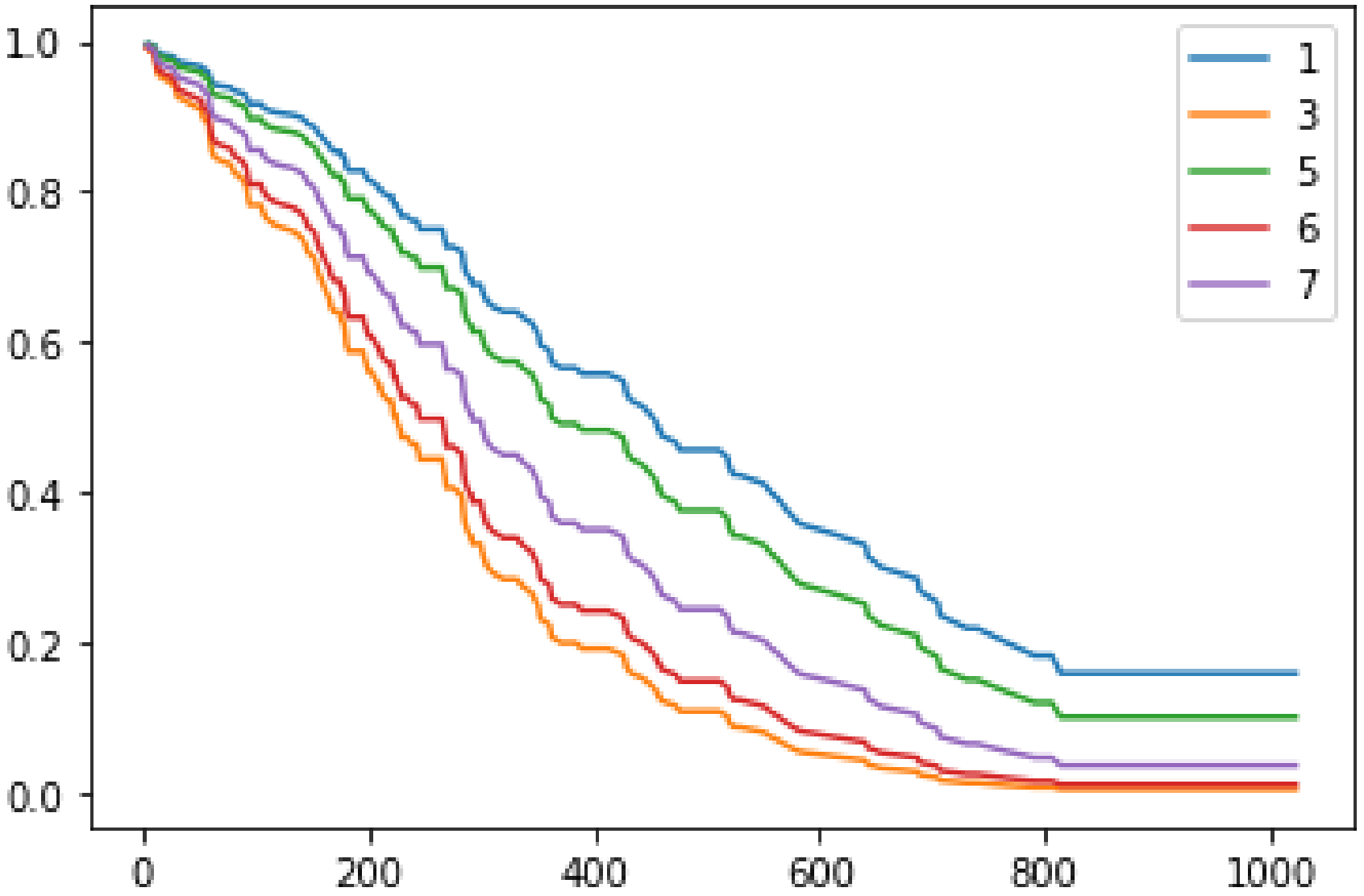
For example, we can say that,

$$S(2) = S_0 * S_1 * S_2$$

In the next article, we'll implement Kaplan-Meier fitter and Nelson-Aalen fitter using python.

Final Result

At the end of this three-part series, you'll be able to plot graphs like this from which we can extrapolate on the survival of a patient. **Hang tight!**



The whole series:

- A Complete Guide To Survival Analysis In Python, part 1
  This three-part series covers a review with step-by-step explanations and code for how to perform statistical survival analysis used to investigate the time some event takes to occur, such as patient survival during the COVID-19 pandemic, the time to failure of engineering products, or even the time to closing a sale after an initial customer contact.
- A Complete Guide To Survival Analysis In Python, part 2
  We look at a detailed example implementing the Kaplan-Meier fitter theory as well as the Nelson-Aalen fitter theory, both with examples and shared code.
- A Complete Guide To Survival Analysis In Python, part 3
  We look at a detailed example implementing the Kaplan-Meier fitter based on different groups, a Log-Rank test, and Cox Regression, all with examples and shared code.

Original. Reposted with permission.

Bio: Pratik Shukla is an aspiring machine learning engineer who loves to put complex theories in simple ways. Pratik pursued his undergraduate in computer science and is going for a master's program in computer science at University of Southern California. “Shoot for the moon. Even if you miss it you will land among the stars. -- Les Brown”

Related:

- Survival Analysis for Business Analytics
- The 8 Basic Statistics Concepts for Data Science
- The Challenges of Building a Predictive Churn Model

What do you think?

29 Responses

Upvote Funny Love Surprised Angry Sad

[<= Previous post]

[Next post =>]

Top Stories Past 30 Days

Most Popular

1. **How I Tripled My Income With Data Science in 18 Months**
2. **How to Build Strong Data Science Portfolio as a Beginner**
3. **What Google Recommends You do Before Taking Their Machine Learning or Data Science Course**
4. **Data Scientist vs Data Engineer Salary**
5. **The 20 Python Packages You Need For Machine Learning and Data Science**

Most Shared

1. **Data Science Portfolio Project Ideas That Can Get You Hired (Or Not)**
2. **Exclusive: OpenAI summarizes KDnuggets**
3. **How I Tripled My Income With Data Science in 18 Months**
4. **Machine Learning Model Development and Model Operations: Principles and Practices**
5. **38 Free Courses on Coursera for Data Science**

Latest News

- Toloka 101 Live Demo: Learn how to get reliable trainin...
- A First Principles Theory of Generalization
- AI Infinite Training & Maintaining Loop
- NLP for Business in the Time of BERTera: Seven Misplace...
- 7 of The Coolest Machine Learning Topics of 2021 at ODS...
- Visual Scoring Techniques for Classification Models

Top Stories
Last Week

Most Popular

1. **How I Tripled My Income With Data Science in 18 Months**
2. **What Google Recommends You do Before Taking Their Machine Learning or Data Science Course**
3. **Learn To Reproduce Papers: Beginner's Guide**
4. **365 Data Science courses free until 18 November**
5. **A Guide to 14 Different Data Science Jobs**

Most Shared

1. **Machine Learning Model Development and Model Operations: Principles and Practices**
2. **A Guide to 14 Different Data Science Jobs**
3. **What Google Recommends You do Before Taking Their Machine Learning or Data Science Course**
4. **Learn To Reproduce Papers: Beginner's Guide**
5. **Four Basic Steps in Data Preparation**

More Recent Stories

- Visual Scoring Techniques for Classification Models
- Data Scientist Career Path from Novice to First Job
- Neural Networks from a Bayesian Perspective
- KDnuggets 21:n42, Nov 3: Google Recommendations Before Taki...
- Three reasons to self-host your product analytics
- ORDAINED: The Python Project Template
- Design Patterns for Machine Learning Pipelines
- Salary Breakdown of the Top Data Science Jobs
- Top Stories, Oct 25-31: How I Tripled My Income With Data Scie...
- Advanced PyTorch Lightning with TorchMetrics and Lightning Flash
- Top 5 Time Series Methods
- Is the Modern Data Stack Leaving You Behind?
- The Case for a Global Responsible AI Framework
- Multivariate Time Series Analysis with an LSTM based RNN
- ETL and ELT: A Guide and Market Analysis
- Simple Text Scraping, Parsing, and Processing with this Python...
- What Google Recommends You do Before Taking Their Machine Learning or Data Science Course **[Silver Blog]**
- Want to Join a Bank? Everything Data Scientists Need to Know A...
- Analyze Python Code in Jupyter Notebooks
- How to Build Data Frameworks with Open Source Tools to Enhance...

X