Departamento de Ciencia de
la Computación (DC-UCSP)

Maestría en Ciencia de Datos

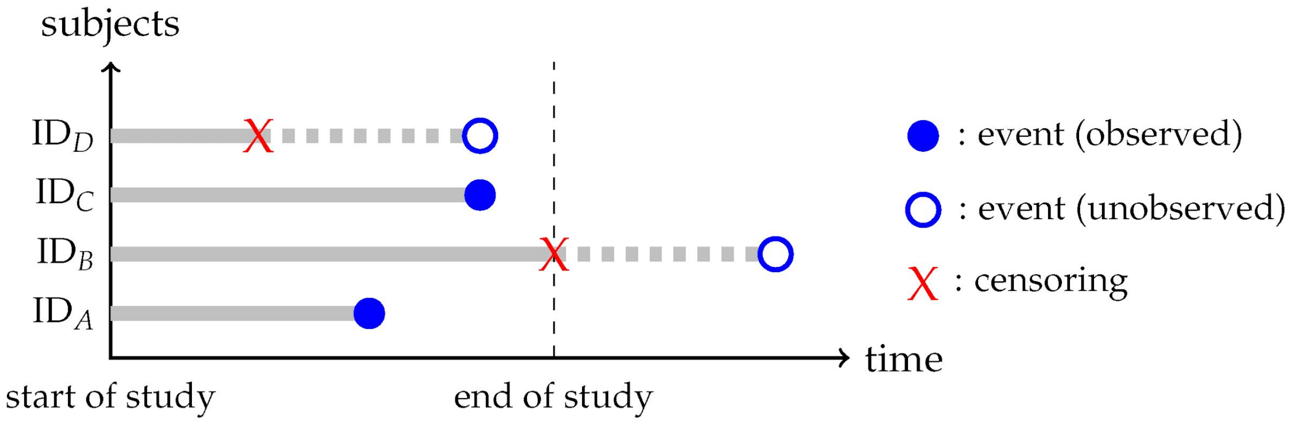
Survival Analysis

Daniel Alexis Gutierrez Pachas, PhD
dgutierrezp@ucsp.edu.pe

Universidad Católica San Pablo | Departamento de Ciencia de la Computación

INTRODUCTION

The analysis of lifetimes is an important topic within biology and medicine in particular but also in reliability analysis with engineering applications. Such data are often highly nonnormally distributed, so that the use of standard linear models is problematic.



Lifetime data are often censored: You do not know the exact lifetime, only that it is longer than a given value. For instance, in a cancer trial, some people are lost to follow-up or simply live beyond the study period. It is an error to ignore the censoring in the statistical analysis, sometimes with extreme consequences. Consider, for instance, the case where a new treatment is introduced towards the end of the study period, so that nearly all the observed lifetimes will be cut short.

MACHINE LEARNING FOR SURVIVAL ANALYSIS: A SURVEY

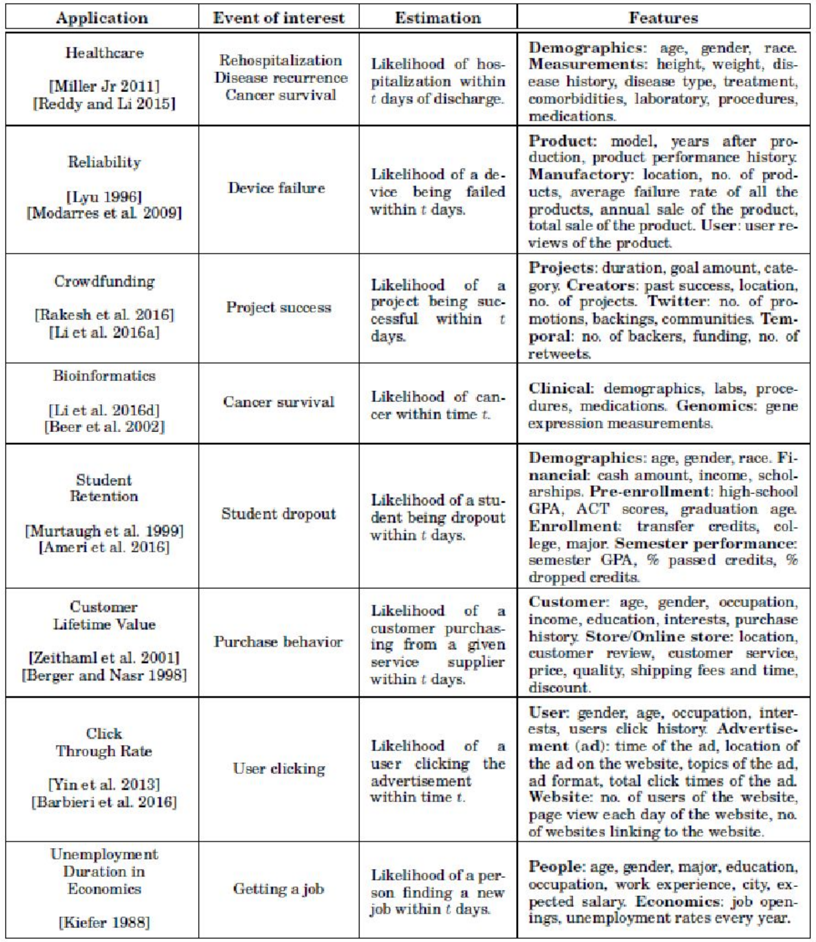
In survival analysis, the outcome variable is “time to an event”. The survival probability function is defined by:

$$S(t)=\text{Prob}(T>t).$$

The hazard function, denoted byh,is defined as the event rate at timetconditional on survivaluntil time t or later (that is,T≤t).

$$h(t) = \lim_{\Delta t \to 0} \frac{\text{Prob}(t \le T \le t + \Delta t \mid t \le T)}{\Delta t}.$$

In addition, H(t)represents the cumulative hazard function.



Application	Event of interest	Estimation	Features
Healthcare [Miller Jr 2011] [Reddy and Li 2015]	Rehospitalization Disease recurrence Cancer survival	Likelihood of hospitalization within t days of discharge.	**Demographics**: age, gender, race. **Measurements**: height, weight, disease history, disease type, treatment, comorbidities, laboratory, procedures, medications.
Reliability [Lyu 1996] [Modarres et al. 2009]	Device failure	Likelihood of a device being failed within t days.	**Product**: model, years after production, product performance history. **Manufactory**: location, no. of products, average failure rate of all the products, annual sale of the product, total sale of the product. **User**: user reviews of the product.
Crowdfunding [Rakesh et al. 2016] [Li et al. 2016a]	Project success	Likelihood of a project being successful within t days.	**Projects**: duration, goal amount, category. **Creators**: past success, location, no. of projects. **Twitter**: no. of promotions, backings, communities. **Temporal**: no. of backers, funding, no. of retweets.
Bioinformatics [Li et al. 2016d] [Beer et al. 2002]	Cancer survival	Likelihood of cancer within time t.	**Clinical**: demographics, labs, procedures, medications. **Genomics**: gene expression measurements.
Student Retention [Murtaugh et al. 1999] [Ameri et al. 2016]	Student dropout	Likelihood of a student being dropout within t days.	**Demographics**: age, gender, race. **Financial**: cash amount, income, scholarships. **Pre-enrollment**: high-school GPA, ACT scores, graduation age. **Enrollment**: transfer credits, college, major. **Semester performance**: semester GPA, % passed credits, % dropped credits.
Customer Lifetime Value [Zeithaml et al. 2001] [Berger and Nasr 1998]	Purchase behavior	Likelihood of a customer purchasing from a given service supplier within t days.	**Customer**: age, gender, occupation, income, education, interests, purchase history. **Store/Online store**: location, customer review, customer service, price, quality, shipping fees and time, discount.
Click Through Rate [Yin et al. 2013] [Barbieri et al. 2016]	User clicking	Likelihood of a user clicking the advertisement within time t.	**User**: gender, age, occupation, interests, users click history. **Advertisement (ad)**: time of the ad, location of the ad on the website, topics of the ad, ad format, total click times of the ad. **Website**: no. of users of the website, page view each day of the website, no. of websites linking to the website.
Unemployment Duration in Economics [Kiefer 1988]	Getting a job	Likelihood of a person finding a new job within t days.	**People**: age, gender, major, education, occupation, work experience, city, expected salary. **Economics**: job openings, unemployment rates every year.

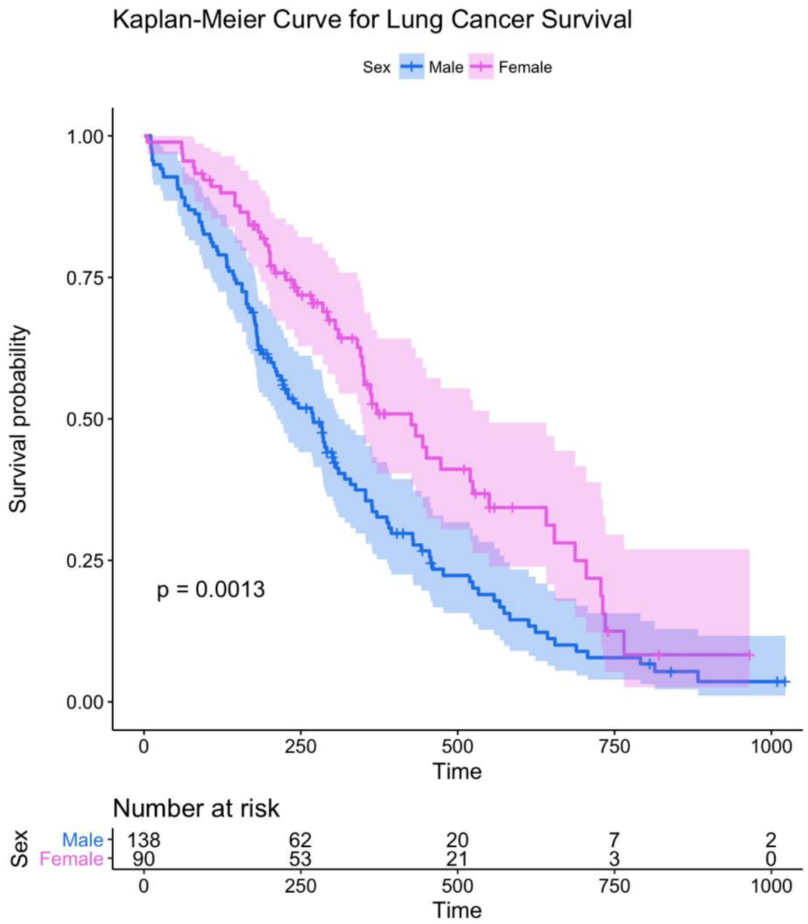
KAPLAN MEIER ESTIMATOR

In 1958, Kaplan and Meier developed the Kaplan-Meier Curve to estimate the survival function using the actual length of the observed time.

$$\hat{S}(t) = \prod_{i:\, t_i \le t} \left(1 - \frac{d_i}{n_i}\right),$$

where d_i is the number of events and $n_i$are the individuals known to have survived. The log-rank test is a hypothesis test to compare the survival distributions of two samples (sub groups A and B). In this case, the null hypothesis is given by:

$$H_0 \colon h_A = h_B.$$



COX REGRESION

Cox Regression builds a predictive model for time-to-event data. The model produces a survival function that predicts the probability that the event of interest has occurred at a given time *t* for given values of the predictor variables. Considering five predictor variables X=(X_1, X_2, X_3, X_4, X_5) and the model

$$\text{h(t|X)} = h_0(t)e^{\beta_1 X_1+\beta_2 X_2+\beta_3 X_3+\beta_4 X_4+\beta_5 X_5}$$

$h_0(t)$ is the baseline hazard function. These methods are different from typical regression/classification because it depends on T. In survival analysis, the measure of effect typically obtained is called a hazard ratio (HR). For example HR of the variable X_3 is:

$$HR_3 = \frac{h(t\,|X_1, X_2, X_3+1, X_4, X_5)}{h(t\,|X_1, X_2, X_3, X_4, X_5)} = e^{\beta_3}.$$

INTERPRETATION OF HAZARD RATIO

Hazard Ratio = Hazard in the intervention group ÷ Hazard in the control group.

Hazard represents the instantaneous event rate, which means the probability that an individual would experience an event at a particular given point in time after the intervention, assuming that this individual has survived to that particular point of time without experiencing any event. Then

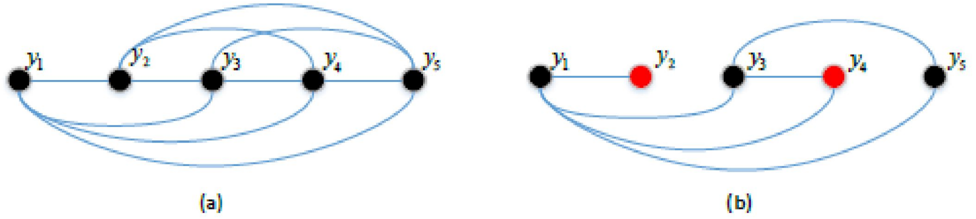
- **HR = 0.5:** at any particular time, **half** as many patients in the treatment group are experiencing an event compared to the control group.
- **HR = 1:** at any particular time, event rates are the **same** in both groups.
- **HR = 2:** at any particular time, **twice** as many patients in the treatment group are experiencing an event compared to the control group.

CONCORDANCE INDEX

In survival analysis, a common way to evaluate a model is to consider the relative risk of an event for different instance instead of the absolute survival times for each instance. This can be done by computing the concordance probability or the concordance index (or C-index). The survival times of two instances can be ordered for two scenarios:

- Both of them are uncensored.
- The observed event time of the uncensored instance is smaller than the censoring time of the censored instance.

By this definition, for the binary prediction problem, C-index will have a similar meaning
to the regular area under the ROC curve (AUC).



(a)

(b)