A short course on Survival Analysis applied to the Financial Industry

BBVA Data & Analytics, Madrid

Marta Sestelo

27-28 /11/2017, v1.0

Contents

Preface

This book is designed to provide a guide for a short course on survival analysis. It is mainly focussed on applying the stastical tecnquines developed in the survival field to the financial industry. The emphasis is placed in understanding the methods, building intuition about when aplying each of them and showing their application through the use of statistical software.

Programing language and software

The **software** used in the course is the statistical language `R`[1] and the IDE (Integrated Development Environment) used is `RStudio`[2]. A basic prior knowledge of both is assumed. Basic introductions to `R` and `RStudio` are presented in the Appendix **??** and **??** for those students lacking basic expertise on them.

The required packages for the course are:

```
# Install packages
install.packages(c("survival", "condSURV", "JM", "dplyr", "survminer", "ggplot2"))
devtools::install_github("noramvillanueva/clustcurv")
```

The codes in the notes may assume that the packages have been loaded, so it is better to do it now:

```
# Load packages
library(survival)
library(condSURV)
library(JM)
library(dplyr)
library(survminer)
library(clustcurv)
```

Links: `survival`[3](**?**), `condSURV`[4](**??**), `JM`[5](**?**), `dplyr`[6](**?**), `survminer`[7](**?**), `ggplot2`[8](**?**), and `clustcurv`[9].

[1] https://cran.r-project.org/
[2] https://www.rstudio.com/products/rstudio/download/
[3] https://cran.r-project.org/web/packages/survival/index.html
[4] https://cran.r-project.org/web/packages/condSURV/index.html
[5] https://cran.r-project.org/web/packages/JM/index.html
[6] https://cran.r-project.org/web/packages/dplyr/index.html
[7] https://cran.r-project.org/web/packages/survminer/index.html
[8] https://cran.r-project.org/web/packages/ggplot2/index.html
[9] https://github.com/noramvillanueva/clustcurv

Main references and credits

Several reference books have been used for preparing these notes. The following list details the most important ones:

- **?**
- **?**
- **?**
- **?**
- **?**
- **?**

In addition, this material is possible due to the work of persons who contribute greatly to the open source software with incredible pieces of software: **?**, **?**, **?** and **?**.

The icons used in the notes were designed by Gregor Cresnar[10] from Flaticon[11].

All material in these notes is licensed under CC BY-NC-SA 4.0[12].

[10] https://www.flaticon.com/authors/gregor-cresna
[11] http://www.flaticon.com/
[12] https://creativecommons.org/licenses/by-nc-sa/4.0/

About the Author

Marta Sestelo is a PhD in Statistics and Data Scientist at the Centre of Mathematics (CMAT) of the University of Minho. She is focused on the development of new methodologies and algorithms linked with statistics. She is interested in estimation and inference methods of flexible models, in developing practical tools for data analysis and in gaining better understanding of real life issues through statistical knowledge. At the moment, her lines of research are closely related to computational statistics and machine learning, particularly, software development, feature selection, applied predictive modeling, nonparametric curves estimation, survival, clustering, model performance, testing procedures, bootstrap resampling methods and applications to different areas of knowledge.

You can see some topics of her cv at `http://sestelo.github.io`.

Introduction

This introduction to survival analysis tries to give a small overview of the statistical approach called survival analysis. This approach includes the type of problem addressed by survival analysis, the outcome variable considered, the need to take into account *censored data*, what a survival function and a hazard function represent, the goals of survival analysis, and some examples of survival analysis.

0.1 What is survival analysis?

In a general way, survival analysis is a collection of statistical procedures for data analysis for which the outcome variable of interest is **time until an event occurs**, often referred to as a failure time, survival time, or event time.

Survival time refers to a variable which measures the time from a particular starting time (e.g., time initiated the treatment) to a particular endpoint of interest: **time-to-event**.

The problem of analyzing time to event data arises in a number of applied fields, such as:

- medicine, biology, public health (time to death)
- social sciences (time for doing some task)
- economics (time looking for employment)
- financial or credit scoring (time to default)
- engineering (time to a failure of some electronic component)

0.1.1 Time, time origen, time scale, event

In survival analysis three requirements are needed for the precise definition of the failure time of an individual. A **time origin** must be specified, a **time scale** for measuring time must be agreed upon and the meaning of **failure - event** must be clear.

- By **time**, we mean years, months, weeks, or days from the beginning of follow-up of an individual until the event of study occurs, but we need to specify the scale.
- By **time origin** we understand the time of entry into the study.
- By **event**, we mean –it depends on the field– death, disease incidence, recovery (e.g., return to work) if we focus on biomedical applications, default in the credit scoring field, renewals in insurance framework, fault in the engeniering field, etc.

Generally, we will assume that only **one event** is of designated interest. When more than one event is considered (e.g., death from any of several causes), the statistical problem can be characterized as either a **recurrent event** or a **competing risk**. We will see the case of the recurrence event using the condSURV[13] package in the Chapter **??**.

[13] `https://cran.r-project.org/web/packages/condSURV/index.html`

It is time to see now an example in a real dataset. This is the **Prosper Loan data** provided by Udacity Data Analyst Nanodegree (last updated 3/11/14). It is also at Kaggle[14].

Prosper.com[15] is a peer-to-peer lending marketplace. Borrowers make loan requests and investors contribute as little as $25 towards the loans of their choice. Historically, Prosper made their loan data public nightly, however, effective January 2015, information will be made available 45 days after the end of the quarter.

A link to the data is here[16] and a variable dictionary can be found here[17].

```
# Prosper Loan data
web <- "https://s3.amazonaws.com/udacity-hosted-downloads/ud651/prosperLoanData.csv"
loan <- read.csv(web)
head(loan)[, c(51, 65, 6, 7, 19, 18, 50)]
##                   LoanKey LoanOriginationDate LoanStatus
## 1 E33A3400205839220442E84 2007-09-12 00:00:00  Completed
## 2 9E3B37071505919926B1D82 2014-03-03 00:00:00    Current
## 3 6954337960046817851BCB2 2007-01-17 00:00:00  Completed
## 4 A03936644658862956 19C51 2012-11-01 00:00:00    Current
## 5 A18036930218888920068 9E 2013-09-20 00:00:00    Current
## 6 C3D63702273952547E79520 2013-12-24 00:00:00    Current
##            ClosedDate    Occupation BorrowerState StatedMonthlyIncome
## 1 2009-08-14 00:00:00         Other            CO            3083.333
## 2                       Professional            CO            6125.000
## 3 2009-12-17 00:00:00         Other            GA            2083.333
## 4                      Skilled Labor            GA            2875.000
## 5                          Executive            MN            9583.333
## 6                       Professional            NM            8333.333
```

0.1.2 Goals of the survival analysis

- Estimate time-to-event for a group of individuals, such as time until default for a group of clients.
- Compare time-to-event between two or more groups, such as residence place for clients.
- Assess the relationship of covariates to time-to-event, such as: occupation, state, income, etc.

0.2 Censoring

The distinguishing feature of survival analysis is that it incorporates a phenomen called **censoring**. Censoring occurs when we have some information about individual survival time, but we don't know the time exactly.

There are generally several reasons why censoring may occur:

- a person does not experience the event before the study ends
- a person is lost to follow-up during the study period
- a person withdraws from the study because of death (if death is not the event of interest) or some other reason
- a person cancela anticipadamente el credito (in credit scoring)

[14] https://www.kaggle.com/jschnessl/prosperloans
[15] https://www.prosper.com/
[16] https://s3.amazonaws.com/udacity-hosted-downloads/ud651/prosperLoanData.csv
[17] https://docs.google.com/spreadsheets/d/1gDyi_L4UvIrLTEC6Wri5nbaMmkGmLQBk-Yx3z0XDEtI/edit#gid=0

- el palzo del credito es inferior a la longitud del estudio y por lo tanto el acreditado cumple integramente con el pago de la deuda antes de concluir el estudio (in credit scoring)

There are three types:

- **Right censoring**: Random right censoring arise often in medical, biological and financial applications. In this studies, patients may enter the study at different times and **the real event time is greater than the observed time**. We know that the person's true survival time becomes incomplete at the right side of the follow-up period, occurring when the study ends or when the person is lost to follow-up or is withdrawn. For these data, the complete survival time interval, which we don't really know, has been cut off (i.e., censored) at the right side of the observed survival time interval. **This is the assumed censoring in the case of credit scoring**.
- **Left censoring**: The survival time of some subject is considered to be left censored if it is less than the value observed. That is, **the event of interest has already occurred for the individual before the observed time** (not easy to deal with). For example, if we are following persons until they become HIV positive, we may record a failure when a subject first tests positive for the virus. However, we may not know the exact time of first exposure to the virus, and therefore do not know exactly when the failure occurred. Thus, the survival time is censored on the left side since the true survival time, which ends at exposure, is shorter than the follow-up time, which ends when the subject's test is positive.
- **Interval censoring**: When **the survival time is only known to occur within an interval**. Such interval censoring occurs when patients in a clinical trial or longitudinal study have periodic follow-up and the patient's event time is only known to fall in some interval. As an example, again considering HIV, a subject may have had two HIV tests, where he/she was HIV negative at the time (say, t_1) of the first test and HIV positive at the time (t_2) of the second test. In such a case, the subject's true survival time occurred after time t_1 and before time t_2, i.e., the subject is interval-censored in the time interval (t_1, t_2).

It is important to highlight in this context (**time-to-default**) which situations we are going to considered as censoring. The bank has special characteristics that are not seen in other applications. **Censored cases** are considered to be loans that did **not experience default** by the moment of data gathering. Additionally, early **repayment** and **mature** cases (or complete, those ones who reach their predefined end date before the moment of data gathering) are also marked censored.

Another classification:

- **Random type I censoring**: Also known as *Generalized Type I Censoring*. When individuals enter the study at different times and the terminal point of the study is predetermined by the investigator, so that the censoring times are known when an individual is entered into the study.
- **Type II censoring**: The study continues until the failure of the first r individuals, where r is some predetermined integer ($r < n$). All subjects are put on test at the same time, and the test is terminated when r of the n subjects have "failed".

0.3 Some notation

We are now ready to introduce **basic mathematical terminology** and **notation** for survival analysis.

Let T the random variable that denotes the survival time, i.e., the time to an event. Since T denotes time, its possible values include all nonnegative numbers; that is, T can be any number equal to or greater than zero. Furthermore, t will be any specific value of interest for the random variable T.

Additionally, when each subject has a random right censoring time C_i that is independent of their failure time T_i, the data is represented by (Y_i, Δ_i) where $Y_i = \min(T_i, C_i)$ and $\Delta_i = I(T_i \leq C_i)$, this Δ define a

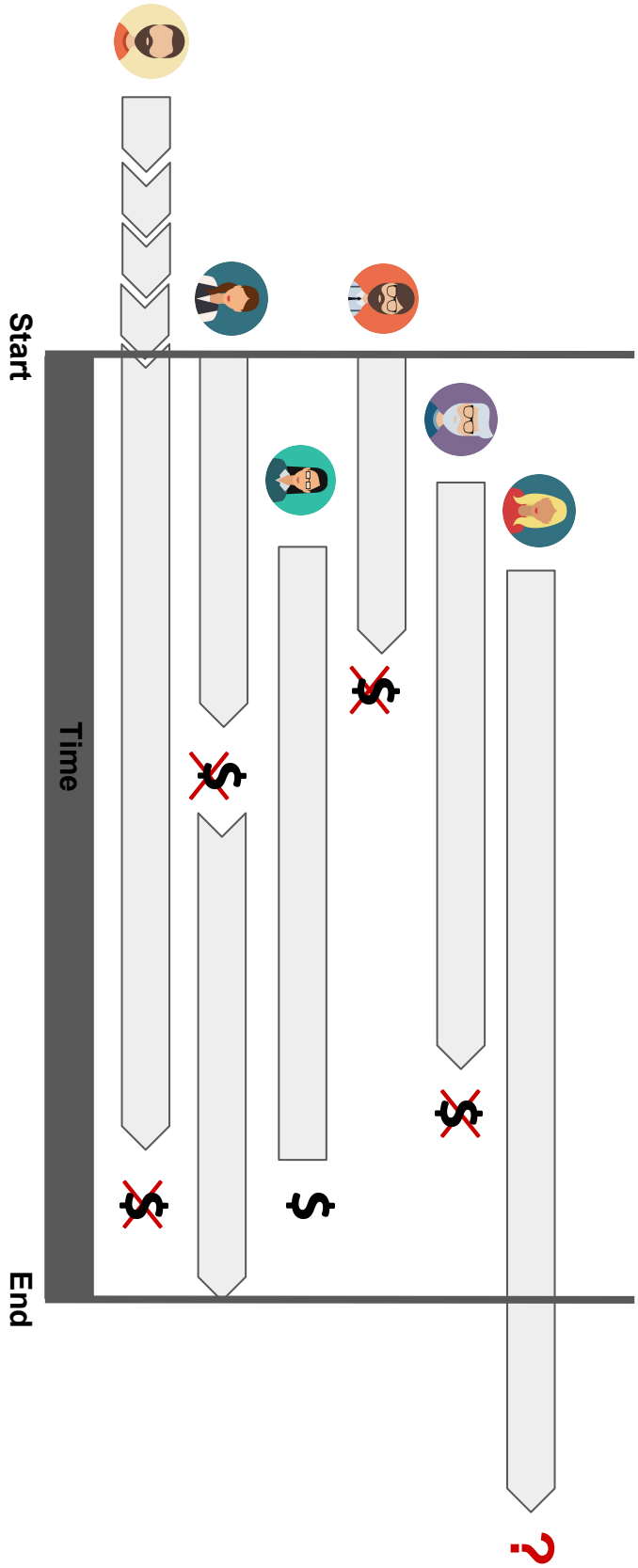


Figure 1: Illustration of censoring.

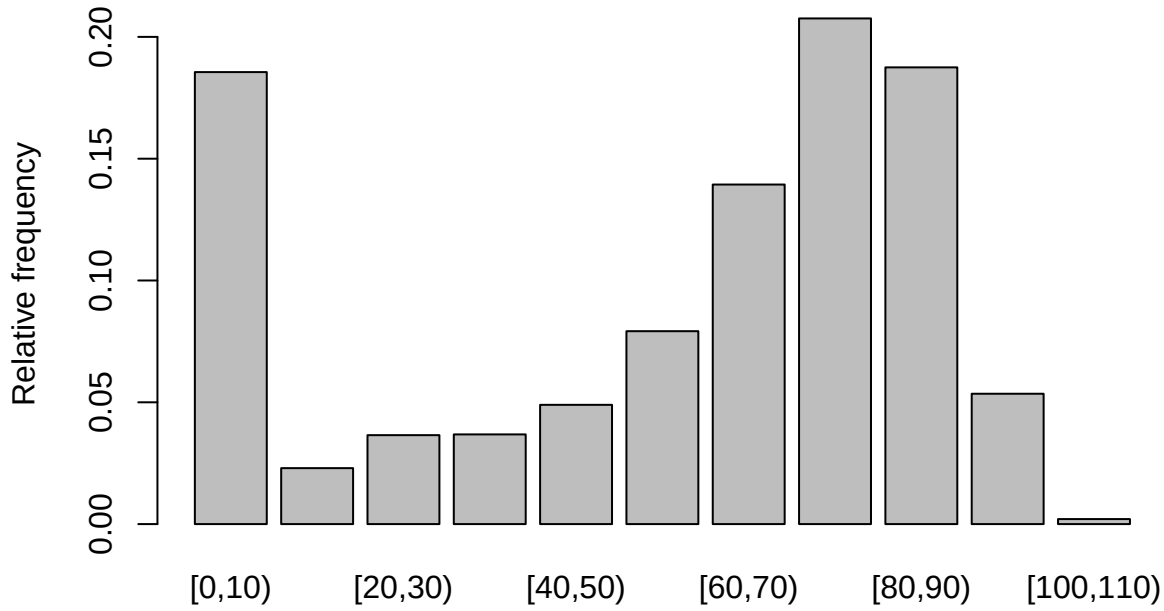


Figure 2: Relative frequiencies for grouped ages.

$(0, 1)$ random variable indicating either failure or censorship. That is, $\Delta = 1$ for failure if the event occurs during the study period, or $\Delta = 0$ if the survival time is censored by the end of the study period.

0.4 Survival/hazard functions

Assuming that T is a continuous non-negative random variable which denote the time-to-event. There is a certain probability that an individual will have an event at exactly time t. For example, about human longevity, human beings have a certain probability of dying at ages 2, 20, 80, and 140, that could be: $P(T = 2)$, $P(T = 20)$, $P(T = 80)$ and $P(T = 140)$.

Similarly, human beings have a certain probability of being alive at those same ages: $P(T > 2)$, $P(T > 20)$, $P(T > 80)$, and $P(T > 140)$.

Here an example with same real data [18]:

```
data <- read.table("data/deaths_esp.txt", header = TRUE, sep = "")
data <- data[!data$Age == "110+", ] # to avoid errors
data$Age_cut <- cut(as.numeric(as.character(data$Age)),
                    breaks =  seq(0,110, 10), right = FALSE)

by_age <- data %>%
  group_by(Age_cut)  %>%
  summarise (sum_deaths = sum(Total, na.rm = TRUE))

barplot(by_age$sum_deaths/sum(data$Total), names.arg = by_age$Age_cut, ylab= "Relative frequency")
```

In the case of human longevity, the probability of death is higher at the beginning and end of life (in Spain). Therefore, T is unlikely to follow a normal distribution. We can see a higher chance of dying (the event of

[18] Data from *The Human Mortality Database* at `http://www.mortality.org`.

interest) in their 70's and 80's and smaller chance of dying in their 100's and 110's, because few people make it long enough to die at these age.

The function that gives the probability of the failure time occurring at exactly time t is the **density function** $f(t)$ [19]

$$f(t) = lim_{\Delta_t \to 0} \frac{P(t \leq T < t + \Delta t)}{\Delta t}$$

and the function that gives the probability of the failure time occur before or exactly at time t is the **cumulative distribution function** $F(t)$

$$F(t) = P(T \leq t) = \int_0^t f(u)du.$$

Note that $F(t)$ is more interesting than $f(t)$... And why? Well, as we said, the main goal of survival analysis is to estimate and compare survival experiences of different groups and the survival experience is described by the **survival function** $S(t)$

$$S(t) = P(T > t) = 1 - F(t)$$

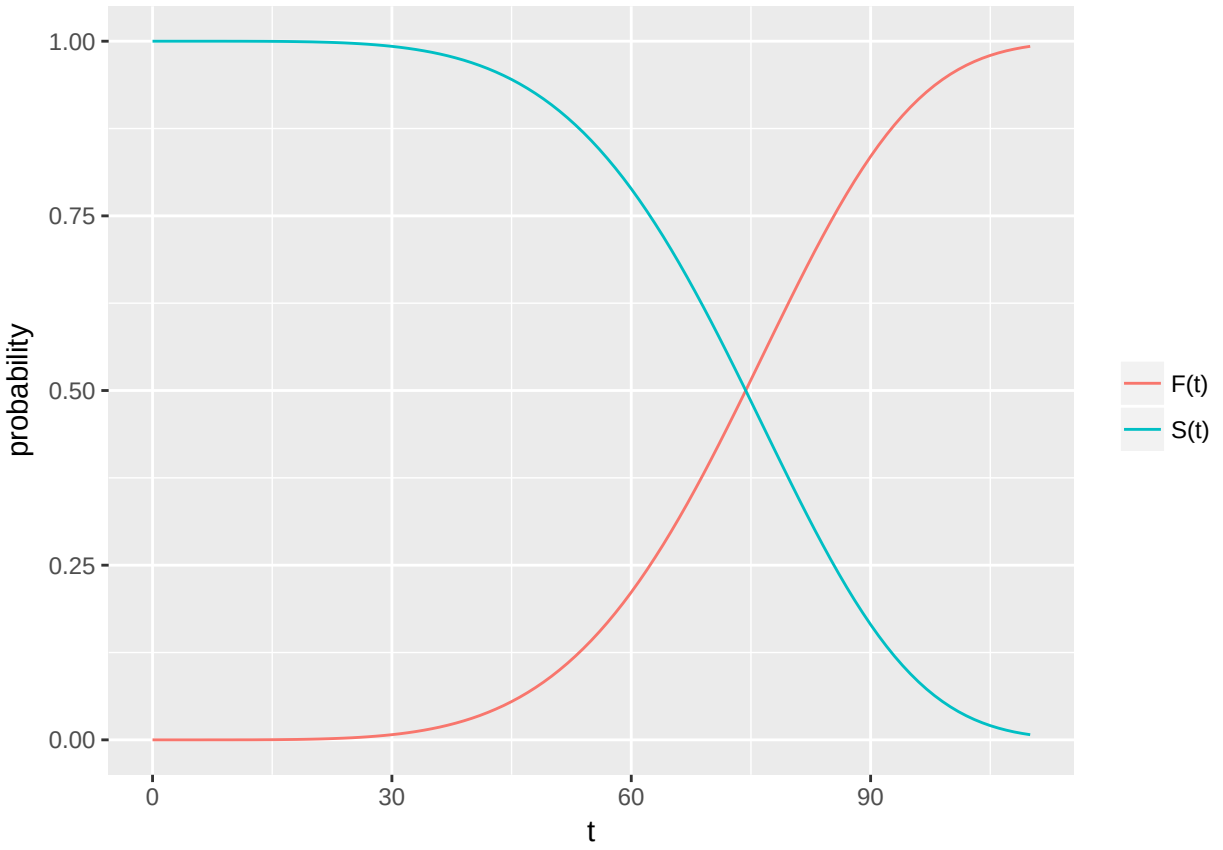
The survival function gives the probability that a person survives longer than some specified time t: that is, $S(t)$ gives the probability that the random variable T exceeds the specified time t. And here, some important characteristics:

- It is nonincreasing; that is, it heads downward as t increases.
- At time $t = 0$, $S(t) = S(0) = 1$; that is, at the start of the study, since no one has gotten the event yet, the probability of surviving past time zero is one.
- At time $t = \text{inf}$, $S(t) = S(\text{inf}) = 0$; that is, theoretically, if the study period increased without limit, eventually nobody would survive, so the survival curve must eventually fall to zero.

```
t <- seq(0, 110, 1)
tdf <- pweibull(t, scale = 80, shape = 5) # weibull dist

d <- reshape2::melt(data.frame(x = t, dist = tdf, surv = 1 - tdf), id = "x")
qplot(x = x, y = value, col = variable, data = d, geom = "line",
      ylab = "probability", xlab = "t") +
  scale_colour_discrete(labels= c("F(t)", "S(t)"), name = "")
```

[19] The probability mass function is a function that gives the probability that a discrete random variable is exactly equal to some value.



Note that these are theoretical properties of survival curves. In practice, when using actual data, we usually obtain graphs that are step functions, rather than smooth curves. Moreover, because the study period is never infinite in length and there may be competing risks for failure, it is possible that not everyone studied gets the event. The estimated survival function, $\hat{S}(t)$ thus may not go all the way down to zero at the end of the study.

```
by_age <- data %>%
  group_by(Age)  %>%
  summarise (sum_deaths = sum(Total, na.rm = T))
t <- rep(as.numeric(as.character(by_age$Age)), by_age$sum_deaths) # real times

aux <- ecdf(t)
x <- seq(0, 110, 1)
edf <- aux(x) # evaluating the ecdf in some points
esf <- 1- edf

d <- reshape2::melt(data.frame(x = x, dist = edf, surv = esf), id = "x")
qplot(x = x, y = value, col = variable, data = d, geom = "step",
      ylab = "Probability", xlab = "t") + scale_colour_discrete(labels = c("F(t)", "S(t)"), name = "")
```