PONTIFICIA UNIVERSIDAD CATÓLICA DEL PERÚ

Postgraduate School

AN APPLICATION OF DISCRETE TIME SURVIVAL MODELS TO ANALYZE STUDENT DROPOUTS AT A PRIVATE UNIVERSITY IN PERÚ

Thesis submitted in partial fulfillment of the requirements for the degree of
Master in Statistics

By

MIGUEL RAÚL PEBES TRUJILLO

Thesis advisor: Giancarlo Sal y Rosas Ph.D.

Examining Committee Members:
Luis Valdivieso Serrano, Ph.D.
Dr. Cristian Bayes Rodríguez
Giancarlo Sal y Rosas, Ph.D.

Lima, Perú. Julio 2015

Dedication

Thanks “papi Raulito” for having been my cornerstone..
To my loved family.. my treasure..

Gratitude

Once again thank God for the opportunity to be here in this world, and for give me the tools I need to live it ♣. All intangibles behind this thesis, effort, time, sacrifices and good intentions, are part of this magic.

I would like to thank my advisor Giancarlo Sal y Rosas, for being my first professor in this learning process and for his invaluable support and feedback; as well as professors Cristian Bayes, Luis Valdivieso and Elizabeth Doig, for their suggestions to improve the content of this work. I would also like to thank Susy Dumett for her generous support in the initial stage of this thesis.

Abstract

Discrete-time survival models are discussed and applied to the study of which factors are associated with student dropouts at a private university in Lima, Perú. We studied the characteristics of $26,790$ incoming students enrolled between 2004 and 2012 in all the undergraduate programs at the University. The analysis include the estimation of the survival and hazard functions using the Kaplan-Meier method and the fitting of parametric models using the Cox proportional hazards regression and the Logistic regression for survival analysis, this last one, in order to include time varying variables as predictors. During the period of analysis, the cumulative probability of remain at the University after five years was 73.7% [95% CI: 73.1% - 74.4%]. In any period the hazard is greater than 4.4% and this highest value is reached in the 3rd semester. In a multivariate analysis, we found that academic factors (area of study, type of admission, standardized academic performance index, and the percentage of passed credits); economic factors (type of residence, and payment scale); and sociodemographic factors (mother education level, indicators of whether or not parents are alive, and the age of the student) were associated with the risk of dropout.

Keywords: survival analysis, discrete survival analysis, time failure analysis, discrete time model, Kaplan−Meier estimator, Cox regression model, logistic model, proportional hazard model, university dropouts.

Contents

List of Abbreviations and Symbols vii

List of Figures viii

List of Tables ix

1 Introduction 1

2 Discrete Time Survival Model 4
- 2.1 Definitions . . . 4
- 2.2 Inference . . . 6
  - 2.2.1 Data structure . . . 6
  - 2.2.2 One sample estimation . . . 8
  - 2.2.3 Regression models . . . 11
  - 2.2.4 Estimation with covariates . . . 17
  - 2.2.5 Particulars for Cox regression . . . 19
  - 2.2.6 Particulars for logistic regression . . . 20
  - 2.2.7 Hypothesis testing . . . 22
  - 2.2.8 Comparison of models with the likelihood test . . . 23

3 Application 25
- 3.1 Framework of the study . . . 25
- 3.2 Data Structure . . . 25
- 3.3 Results . . . 26

4 Discussion and Future Research 38
- 4.1 Conclusions . . . 38
- 4.2 Future Research at Survival Analysis . . . 39

A Theoretical notes and calculations 40
  - A.0.1 Probability density function of T expressed in terms of the hazard function . . . 40
  - A.0.2 Hazard ratio interpretation . . . 40
  - A.0.3 Proportional hazard assumption . . . 41
  - A.0.4 Likelihood ratio test . . . 42

B Graphical results **43**

B.0.5 Kaplan-Meier estimators for the whole population 43

B.0.6 Estimated hazard functions or logit(hazard) functions for relevant covariates 43

C Source code **45**

C.0.7 Simulation of proportionality assumption 45

C.0.8 Processing source code in R 3.0 45

Bibliography **55**

List of Abbreviations and Symbols

HR	Hazard ratio.
OR	Odds ratio.
SE	Standard error.
ML	Maximum likelihood.
MLE	Maximum likelihood estimator.
PL	Partial likelihood.
e or exp	Exponential.
I	Indicator function.
cloglog	Complementary log-log.
CI	Confidence interval.
E	Expected value (mean).
Var	Variance.
σ^2	Variance.
σ	Standard deviation.
se	Standard error.
Ref	Reference.
DF	Degrees of freedom.
χ^2	Chi-square.
$\lambda(x)$	Hazard function of x.
$\Lambda(x)$	Cumulative hazard function of x.
$S(x)$	Survival function of x.
$F(x)$	Cumulative distribution function of x.
$f(x)$	Probability density function of x.

List of Figures

2.1 Continuous and discrete intervals 5
2.2 Mechanisms of censoring 7
2.3 Proportionality hazard assumption in the Cox model: In this example two hazard functions (for two types of individuals) are shown, they are proportional along time. 13
2.4 Proportionality Logit assumption in the logistic model. In this example two Logit(hazard) functions (for two types of individuals) are shown, they are proportional along time. 15

B.1 Kaplan-Meier estimator for the entire population 43
B.2 Hazard for entire population by first enrolled area 44
B.3 Hazard for entire population by residence type 44
B.4 Hazard for entire population by mother level education 44
B.5 Hazard for entire population by admission type 44

List of Tables

2.1 Mapping the failure time as time indicators 16
2.2 "Individual level" example data set 21
2.3 "Individual-period level" example data set 21

3.1 Descriptive statistics at enrollment 27
3.2 Descriptive statistics at enrollment 28
3.3 Descriptive statistics at enrollment 29
3.4 Kaplan-Meier estimation for the entire population 30
3.5 Univariate regression analysis (1/3) 33
3.6 Univariate regression analysis (2/3) 34
3.7 Univariate regression analysis (3/3) 35
3.8 Multivariate logistic regression model (1/2) 36
3.9 Multivariate logistic regression model (2/2) 37

A.1 Likelihood ratio test 42

Chapter 1

Introduction

"I do not want to foresee the future.
I am concerned with taking care of the present.
God has given me no control over the moment following".
Gandhi, 1924

In many areas of science, researchers are interested in estimating the time at which a given event occurs and how this time is accelerated or retarded by certain factors. In this settings, this gathered information of times in which events occur is called time-to-event data. It is often the case when the event of interest can only occur at regular, discrete time points, or the available data is only a fixed interval in which the event occurs. As a motivation example of discrete time, consider the study of student dropout at a given university. This information is only assessed at the end of each semester and the collection process could last up to eight years that is the maximum time a student can remain at the University. As a result, each student event time will be discrete. On the other hand, if we want to estimate the distribution of the student attrition time we would use a random sample of students that could, or not, drop out. Each student that does not drop out can be considered as an incomplete observation because they do not provide a dropout time or never experienced the event of interest. This scenario, as others we are going to explain in the present study, is named "censoring".

Kaplan and Meier (1958) developed the nonparametric estimator of the survival function of the distribution of time to the event. Greenwood (1926) proposed an estimator for the variance and derived approximated confidence intervals for the Kaplan-Meier estimator within an application on health. Cox (1972) proposed the proportional hazards model to continuous times by parameterizing the hazard as a function of explanatory variables and an unknown baseline function of time (baseline hazard function). The initial theoretical development considers a model were failure times are continuous, measured by an exact method and in which two or more individuals can not fail at the same time (without ties). In the same publication, Cox (1972) proposed another model for the hazard using the Logit or "log(odds)" transformation: the logistic regression model, that could be applied when the failure times are intrinsically discrete. Three years later, Cox (1975) introduced the concept of partial likelihood to reduce the dimensionality of the problem and proved that such a estimator is consistent and has normal distribution. Breslow (1974) and Efron (1977) proposed variations of the Cox likelihood function to deal with ties in the estimation process.

There have been many applications of discrete time survival models in several areas, mostly of them using the theoretical background developed by Kaplan and Meier (1958) and Cox (1972). Allison (1982) used a discrete-time model for the analysis of event histories on a sociological context. Singer and Willett (1993) published an study explaining the importance of discrete time models with special emphasis on educational contexts, taking as an example an analysis of the number of years that Michigan special educators teach in public schools. Scheike and Jensen (1997) worked a discrete time analysis to model the time to pregnancy using the number of menstrual cycles that a couple needs to conceive as a measure of human reproduction. Another study in the medical field was done by Borges (2005), studying survival time of patients with peritoneal dialysis. This study modeled the time since the beginning of dialysis sessions until death as event of interest. Muthn and Masyn (2005) described an analysis of discrete time model incorporating categorical and continuous latent variables and showing its application with two examples: recidivism after incarceration and school removal related with aggressive behavior in classrooms.

Another scientific and specific interest has been represented by attempts to model the dropout time of students in academic institutions and to discover which factors are associated with this phenomena. Given that the student dropout is generally recorded in semesters or academic cycles, current research suggests the use of discrete time survival methods. DesJardins et al. (1998) worked in an educational application, modeling student departures from college and describing the particular effect of different explanatory variables over time. In this study they estimated a time−constant coefficients model and a time−varying coefficients model based in the discrete-time model proposed by McCall (1994), which use a complementary log-log link function to parametrize the hazard. Smith and Naylor (2001) conducted an study to model the probability of withdrawal, for a group of universities in UK, applying a binomial probit regression analysis and considering as withdrawal the departure before a given point in the scheduled final year of the degree course. Later, Arulampalam et al. (2004) applied the proportional hazards model to analyze the probability of medical school dropouts in UK. Radcliffe et al. (2006) modeled the incidence and timing of student attrition using a logit and longitudinal model, being the goal of this study to identify at−risk students to promote retention policies. They found that the powerful predictors were related with academic performance. Recently, Cheng (2012) has studied the effect of institutional characteristics over the dropout risk in higher education institutions in US.

In the current study, we applied the Logistic and Cox proportional hazard models for discrete time-to-event data to data collected between March 2002 and December 2012 of students enrolled in the undergraduate programs at a private University in Lima, Perú. The general aim of this paper is to study regression models for discrete-time survival data and to apply them to estimate student dropouts hazards and understand how student features are related to these rates.

The organization of this document is as follows. In Chapter 1, we define the framework for the study, starting with the definition of survival analysis and presenting a state of the art review considering mainly examples and the most current research in discrete−time survival analysis applied to model student dropouts. In Chapter 2, we present the notation to be used along the study and develop the theoretical background of discrete-time survival models. We show the nonparametric and parametric estimations of the model, with special emphasis on the Kaplan-Meier estimator and the regression models (Cox and logistic). In Chapter 3, we apply the studied methods to understand the students dropout problem at a private University in Perú. It includes the application of all the items described in the theoretical chapter. Results are supported by the implementation of a program in the statistical software R. In Chapter 4 we discuss some outcomes and conclusions obtained in this project, related with the most powerful predictors that explain the study problem, and analyze advantages and disadvantages of the proposed methods. In the section Appendix we show some detailed tests and results (Appendix B), as well as the source code of the applications used to process the data and generate the described results (Appendix C).

Chapter 2

Discrete Time Survival Model

"When I admire the wonders of a sunset or the beauty of the moon
my soul expands in the worship of the creator".
Gandhi, 1924
♣ *Mount Abu, Rajasthan; India*

2.1 Definitions

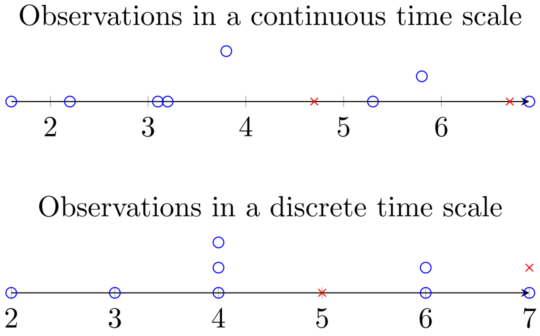
Survival analysis is a group of statistical methods that aims to model the time at which a given event occurs and how this time is accelerated or retarded by certain factors. This time can be continuous or discrete, depending of the scenario of study.

To analyze data in discrete intervals, the continue scale is divided into a infinite sequence of intervals as follows: $[0 = a_0, a_1], (a_1, a_2], \dots, (a_{j-1}, a_j], \dots, (a_{k-1}, a_k = \infty]$, where $a_1, a_2, a_3, \dots, a_j, \dots$ are ordered points in time on which events can occur, j represents an index period, and the jth period is written as $(a_{j-1}, a_j]$. These intervals are usually assumed equal in length (for instance a "month" or a "semester"), in which case positive integers numbers can be used to index, and $a_j = 1, 2, 3, \dots$.

There are two ways in which discrete survival data appear: (i) the survival time is originally continuous but it is only observed in intervals (case known as "grouped data"), or (ii) the survival time scale is originally discrete. In both cases theory assumes that observations are measured in intervals, which means that "failed" observations in different points within a time period have the same survival time (See figure (2.1)).

Let T be a non-negative discrete random variable that represents the time in which the event of interest is observed (failure time). The distribution of T at time "t" can be represented by (i) the hazard function $\lambda(t) = \lambda_t$, (ii) the survival function $S(t) = S_t$, (iii) the cumulative distribution function $F(t) = F_t$, (iv) the probability density function $f(t) = f_t$ and (v) the

Figure 2.1: Continuous and discrete intervals



cumulative hazard function $\Lambda(t) = \Lambda_t$.

The hazard is a fundamental quantity that represents the risk of a event occurrence in a time period. The hazard for an individual i at time $T = j$ is defined as

$$\lambda_{ij} = P(T_i = j \mid T_i \geq j) \tag{2.1}$$

and it is the conditional probability that the event occurs for the individual i at the time period j (from now on only *time* j) given that it did not occur before or, equivalently, given that it has survived until time j. The set of discrete time hazard probabilities, one per period, is known as the discrete time hazard function. In a population the hazard function for any individual can be expressed as $\lambda_j = P(T = j \mid T \geq j)$.

The survival at time j is the probability that an individual survives past time j, which means that the event did not occur neither in a current period nor an earlier time. The set of survival probabilities is known as the survival function. The survival probability that an event has not occur for the individual i at the time period $T = j$, is defined as

$$S_{ij} = P(T_i > j) \tag{2.2}$$

In a population the survival function for any individual can be expressed as $S_j = P(T > j)$. Then

$$\begin{aligned} \lambda_j = P(T = j | T \geq j) &= \frac{P(T = j)}{P(T \geq j)} = \frac{P(T = j)}{P(T = j) + P(T > j)} \\ &= \frac{S_{j-1} - S_j}{[S_{j-1} - S_j] + S_j} = 1 - \frac{S_j}{S_{j-1}} \end{aligned} \tag{2.3}$$

Therefore, the survival function can be expressed as

$$S_j = (1 - \lambda_j) S_{j-1}, \quad j = 1, 2, 3, \ldots \tag{2.4}$$

where $(1 - \lambda_j)$ denotes about the probability of non-occurrence of the event at time j, given that the event did not occur before, and $S_0 = P(T > 0) = 1$, because at the beginning all individuals are assumed to be “alive”, i.e., without experiencing the event.

It is easy to see that the discrete time survival function can be written in terms of the hazard rate as

$$\begin{aligned} S_j &= (1-\lambda_j)(1-\lambda_{j-1})\dots(1-\lambda_2)(1-\lambda_1)(S_0) \\ &= \prod_{k=1}^{j}(1-\lambda_k). \end{aligned} \tag{2.5}$$

Similarly, the discrete time failure cumulative function F_j represents the probability that the event occurs until certain time $T = j$ and is defined as

$$\begin{aligned} F_j = P(T \leq j) &= 1 - S_j \\ &= 1 - \prod_{k=1}^{j}(1-\lambda_k) \end{aligned} \tag{2.6}$$

The probability density function, on the other hand, assesses the probability that the event occurs at time $T = j$ and is represented in terms of the previously defined functions as

$$\begin{aligned} f_j = P(T=j) = F_j - F_{j-1} = S_{j-1} - S_j &= \prod_{k=1}^{j-1}(1-\lambda_k) - \prod_{k=1}^{j}(1-\lambda_k) \\ &= \prod_{k=1}^{j-1}(1-\lambda_k) - (1-\lambda_k)\prod_{k=1}^{j-1}(1-\lambda_k) \\ &= [1-(1-\lambda_j)]\prod_{k=1}^{j-1}(1-\lambda_k) \\ &= \lambda_j \prod_{k=1}^{j-1}(1-\lambda_k) \end{aligned} \tag{2.7}$$

For an alternative demonstration see appendix (A.0.1).

2.2 Inference

2.2.1 Data structure

Let $\{(T_i)\}_{i=1}^{n}$ be a sample of n observations, where T_i is the observed time of the event for the individual i and let j be the index for the intervals in which individuals can fail, where $j = 1, \dots, k$. Then the empirical distribution function at time j, in terms of the indicator function, is defined as

$$\hat{F}_j = \frac{1}{n}\sum_{i=1}^{n} I(T_i \leq j) \tag{2.8}$$

and the empirical survival function is defined as

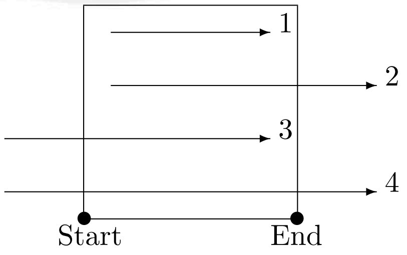
$$\begin{aligned} \hat{S}_j &= 1 - \hat{F}_j \\ &= 1 - \frac{1}{n}\sum_{i=1}^{n} I(T_i \leq j) \end{aligned} \tag{2.9}$$

where $\hat{S}_j$ represents the observed proportion of the n individuals in the population that remain "alive" at $T = j$.

Beyond the natural complexity of modeling the effect of some features that accelerates or retards the time in which an event occurs, event history data has a special feature: censoring. A censored observation is defined as an individual for whom its failure time is unknown, either because it does not experience the event of interest within the observation period (case known as right-censoring), or because at the beginning of the observation its lifetime had already started (case known as left-truncation). Those observations are considered as missing data. Right-censored observations are included in the estimation processes assuming a non informative censoring, which means that the distribution of censoring times is independent of the distribution of the time to the event of interest (this scenario is also known as "missing at random" assumption).

In general, it is possible to identify the following censoring cases when the information about survival time of elements is incomplete. Figure (2.2) shows the types of failure time observations and censoring mechanisms. The arrowheads indicate the time when the event occurs and the whole length of the arrows represents the observation period.

Figure 2.2: Mechanisms of censoring. Arrow 1: Start and end time known. Arrow 2: End time outside the observation period (right-censored observation). Arrow 3: Start time outside the observation period (left-truncated observation). Arrow 4: Start and end time outside the observation period.



Let T_c be the random variable of censoring time and T_0 be the random variable of failure time. Then, one can define the random variable of observed time

$$T = min\{T_0, T_c\}, \tag{2.10}$$

and the censoring indicator

$$C = I(T_0 \leq T_c) = \begin{cases} 1, & \text{when observed time is a failure } (T_0 \leq T_c) \\ 0, & \text{when observed time is censored } (T_0 > T_c) \end{cases} \tag{2.11}$$

2.2.2 One sample estimation

Let $\{(T_i, C_i)\}_{i=1}^{n}$ be a sample of n independent and identically distributed observations, where T_i is the observed time of the event for the individual i and C_i the related censoring indicator.

The contribution to the likelihood function of an individual who fails at time t_i is $P(T = t_i) = f_i$, and the contribution of an censored observation at time t_i is $P(T > t_i) = S_i$ (because we only know that the survival time is greater than t_i). Then, the likelihood function has the form:

$$L(\underline{\lambda} \mid t_1, t_2, \ldots, t_n) = \prod_{i=1}^{n} [P(T = t_i)]^{c_i} [P(T > t_i)]^{1-c_i} \tag{2.12}$$

In the discrete case, there could be more than one failure or censored observation in the same period. Let k be the number of distinct intervals in study on which a failure can occurs, d_j be the number of observations that fail, m_j the number of censored observations, and n_j the number of observations at risk at the jth period, then $n_j = (m_j + d_j) + \ldots + (m_k + d_k)$ (and $n_{j+1} = n_j - d_j - m_j$), where $j = 1, \ldots, k$. Based on this information and according to equations (2.7) and (2.5) the likelihood function L can be written in terms of the hazard as

$$L(\underline{\lambda} \mid t_1, t_2, \ldots, t_n) = \prod_{j=1}^{k} \left\{ \left[\lambda_j \prod_{l=1}^{j-1} (1 - \lambda_l) \right]^{d_j} \left[\prod_{l=1}^{j} (1 - \lambda_l) \right]^{m_j} \right\} \tag{2.13}$$

and analyzing the factor $A_j = \left[\prod_{l=1}^{j-1}(1-\lambda_l)\right]^{d_j} \left[\prod_{l=1}^{j}(1-\lambda_l)\right]^{m_j}$ we can observe the following series of components

$$\begin{array}{lll}
\text{If} \quad j = 1 & , & A_1 = (1 - \lambda_1)^{m_1} \\
\text{If} \quad j = 2 & , & A_2 = (1 - \lambda_1)^{d_2} (1 - \lambda_1)^{m_2} (1 - \lambda_2)^{m_2} \\
\text{If} \quad j = 3 & , & A_3 = (1 - \lambda_1)^{d_3} (1 - \lambda_2)^{d_3} (1 - \lambda_1)^{m_3} (1 - \lambda_2)^{m_3} (1 - \lambda_3)^{m_3} \\
\quad\quad \vdots & & \quad \vdots \\
\text{If} \quad j = k & , & A_k = (1 - \lambda_1)^{d_k} (1 - \lambda_2)^{d_k} \ldots (1 - \lambda_{k-1})^{d_k} (1 - \lambda_1)^{m_k} (1 - \lambda_2)^{m_k} \ldots (1 - \lambda_k)^{m_k}
\end{array} \tag{2.14}$$

Grouping the components with the same λ_l we have

$$\begin{aligned}(1-\lambda_1)^{d_2+\ldots+d_k+m_1+\ldots+m_k} &= (1-\lambda_1)^{n_1-d_1}\\ (1-\lambda_2)^{d_3+\ldots+d_k+m_2+\ldots+m_k} &= (1-\lambda_2)^{n_2-d_2}\\ &\vdots\\ (1-\lambda_k)^{m_k} &= (1-\lambda_k)^{n_k-d_k}\end{aligned}$$

Therefore, L can be reduced to

$$L(\underline{\lambda} \mid t_1, t_2, \ldots, t_n) = \prod_{j=1}^{k} \lambda_j^{d_j}(1-\lambda_j)^{n_j-d_j} \tag{2.15}$$

and the log likelihood function to

$$l(\underline{\lambda} \mid t_1, t_2, \ldots, t_n) = \sum_{j=1}^{k} \left[d_j \log(\lambda_j) + (n_j - d_j)\log(1-\lambda_j)\right]. \tag{2.16}$$

The derivative of the log likelihood function is given by

$$\begin{aligned}\frac{\partial l}{\partial \lambda_i} &= \frac{d_i}{\lambda_i} + \frac{(n_i - d_i)}{(1-\lambda_i)}(-1)\\ &= \frac{d_i - \lambda_i n_i}{\lambda_i(1-\lambda_i)}\end{aligned} \tag{2.17}$$

By equating to zero the derivative of the log likelihood function, it is easy to see that the MLE, $\hat{\underline{\lambda}} = (\hat{\lambda}_1, \ldots, \hat{\lambda}_k)$ of $\underline{\lambda}$, also known as the Kaplan-Meier estimator (Kaplan and Meier (1958)), is defined as

$$\hat{\lambda}_j = \frac{d_j}{n_j}, \qquad j = 1, \ldots, k \tag{2.18}$$

where $\hat{\lambda}_j$ represents the observed proportion of the n_j individuals at risk who fail at $T = j$. Therefore, the estimator for the survival function is defined as

$$\hat{S}_j = \prod_{i=1}^{j}\left(1-\hat{\lambda}_j\right) = \prod_{i=1}^{j}\left(1-\frac{d_i}{n_i}\right) = \prod_{i=1}^{j}\left(\frac{n_i-d_i}{n_i}\right) \tag{2.19}$$

Another important estimator is the Nelson-Aalen estimator of Λ_j, which defines a right-continuous step function approximation of the cumulative risk function based on the Kaplan-Meier estimations, as follows

$$\hat{\Lambda}_j = \sum_{i \mid t_i \le j} \hat{\lambda}_i = \sum_{i \mid t_i \le j} \frac{d_i}{n_i} \tag{2.20}$$

A $100(1-\alpha)\%$ confidence interval for the survival function estimator can be given by considering the Greenwood formula (See Greenwood (1926)). This is based on the first order

Taylor expression

$$f(X) \approx f(c) + f'(c)(X - c), \tag{2.21}$$

of a function $f(X)$ of a random variable X around a point c close to $E(X)$. Then,

$$\begin{aligned} E[f(X)] &\approx f(c) + f'(c)(E(X) - c) \\ Var[f(X)] &\approx f'(c)^2 Var(X). \end{aligned} \tag{2.22}$$

According to equation (2.19) we have

$$log(\hat{S}_j) = \sum_{i=1}^{j} log(\frac{n_i - d_i}{n_i}) = \sum_{i=1}^{j} log(\hat{\pi}_i) \tag{2.23}$$

where $\hat{\pi}_i = \frac{n_i - d_i}{n_i}$. Then applying the variance operator to both sides one can prove

$$Var\Big[log(\hat{S}_j)\Big] = Var\Big[\sum_{i=1}^{j} log(\hat{\pi}_i)\Big] =^{1} \sum_{i=1}^{j} Var\Big[log(\hat{\pi}_i)\Big] \tag{2.24}$$

under the assumption that failures arise independently (independence of variables $\hat{\pi}_i$). Furthermore, if one supposes that d_i follows a Binomial distribution with parameters n_i and π_i, then an estimator of π_i is $\hat{\pi}_i$ and the estimator of the variance is $n_i\hat{\pi}_i(1 - \hat{\pi}_i)$. Given this information and applying variance properties it is easy to see that

$$\begin{aligned} Var(\hat{\pi}_i) &= Var\left(\frac{n_i - d_i}{n_i}\right) = Var\left(1 - \frac{d_i}{n_i}\right) = Var\left(\frac{d_i}{n_i}\right) \\ &= \frac{Var(d_i)}{{n_i}^2} = \frac{n_i\pi_i(1 - \pi_i)}{{n_i}^2} \\ &= \frac{\pi_i(1 - \pi_i)}{n_i} \approx \frac{\hat{\pi}_i(1 - \hat{\pi}_i)}{n_i} \end{aligned} \tag{2.25}$$

Then, considering the last result in equation (2.22) one obtains

$$Var\Big[log(\hat{\pi}_i)\Big] \approx \left(\frac{1}{\pi_i}\right)^2 Var(\hat{\pi}_i) = \frac{1 - \hat{\pi}_i}{n_i\hat{\pi}_i} \tag{2.26}$$

and

$$\begin{aligned} Var\Big[log(\hat{S}_j)\Big] &\approx \sum_{i=1}^{j} Var\Big[log(\hat{\pi}_i)\Big] \\ &\approx \sum_{i=1}^{j} \frac{1 - \hat{\pi}_i}{n_i\hat{\pi}_i} \\ &\approx \sum_{i=1}^{j} \frac{d_i}{n_i(n_i - d_i)}, \hat{\pi}_i = 1 - d_i/n_i \end{aligned} \tag{2.27}$$

[1] This result is not trivial, but is out of the scope to the thesis

A new application of equation (2.22) yields

$$\begin{aligned} \hat{Var}[\hat{S}_j] &= \frac{Var[f(x)]}{f'(x)^2}, f(x) = e^x \\ &= {S_j}^2 \sum_{i=1}^{j} \frac{d_i}{n_i(n_i - d_i)} \end{aligned} \quad (2.28)$$

Let z_α be the α-quantile of the normal distribution. Hence, the $100(1-\alpha)$ % confidence interval of $\hat{S}_j$ is defined as

$$\begin{aligned} \hat{S}_j &\pm z_{\alpha/2}\sqrt{Var[\hat{S}_j]} \quad (2.29) \\ \text{where} \quad \hat{Var}[\hat{S}_j] &= \hat{S}_j^2 \sum_{i=1}^{j} \frac{d_j}{n_j(n_j - d_j)} \end{aligned}$$

observing that $\hat{S} = 1 - \hat{F}$ and

$$\begin{aligned} Var(\hat{S}) &= Var(1 - \hat{F}) \\ &= Var(\hat{F}) \end{aligned}$$

we can also obtain a $100(1-\alpha)$ % confidence interval for $\hat{F}$ given by

$$\hat{F}_j \pm z_{\alpha/2}\sqrt{\hat{Var}[\hat{S}_j]} \quad (2.30)$$

Aalen (1976) and Peterson (1977) demonstrated that the estimator $\hat{S}_j$ has almost sure consistency, furthermore Breslow and J. (1974) established an asymptotic normality of the estimator: $\sqrt{n}(\hat{S}_j - S_j) \rightarrow \mathcal{N}(0, {\sigma^2}_j)$, where σ^2 is estimated as $Var(\hat{S})$.

2.2.3 Regression models

To analyze the predictors that accelerate or delay the failure time we can fit regression models by representing the hazard or an equivalent function as a function of the covariates and evaluate them to analyze its goodness-of-fit. Let $\boldsymbol{z}$ be a set of p factors or explanatory variables for an individual i denoted by $\boldsymbol{z_i} = (z_{1i}, z_{2i}, ..., z_{pi})$, then the hazard at time $T = j$ is denoted by

$$\begin{aligned} \lambda(t_j, \boldsymbol{z}_i) &= P(T = j \mid T \geq j \text{ and } Z = \boldsymbol{z}_i) \\ &= P(T = j \mid T \geq j \text{ and } Z_1 = z_{1i}, Z_2 = z_{2i}, ..., Z_p = z_{pi}) \end{aligned} \quad (2.31)$$

that represents the conditional probability that the event occurs at time t_j given that it did not occur before and given an specific set of variables $\boldsymbol{z}$ for the p predictors. In a heterogeneous population, individuals can have different hazards functions depending of their values for the set of predictors (or profiles) and only if two individuals have the same set of covariate values they will have the same hazard function.

In this work we will consider a regression model for survival analysis of the form

$$y[\lambda(t_j,\boldsymbol{z}_i)] = y[\lambda_0(t_j)] + \boldsymbol{z_i}^t\boldsymbol{\beta} \tag{2.32}$$

or

$$\lambda(t_j,\boldsymbol{z}_i) = y^{-1}[y[\lambda_0(t_j)] + \boldsymbol{z_i}^t\boldsymbol{\beta}] \tag{2.33}$$

where the component $(\boldsymbol{z}^t\beta)$ represents the effect of the predictors, $\boldsymbol{\beta} = (\beta_1, \beta_2, ..., \beta_p)$ is a regression vector of slope parameters, y is a monotone-increasing and twice-differentiable function defined in $(-\infty, \infty)$ with $y[0] = -\infty$, that can be any of the following forms:

- $y(u) = log(u)$, called the discrete relative risk model
- $y(u) = log[u/(1-u)]$, called the discrete logistic model
- $y(u) = log[-log(1-u)]$, called the grouped relative risk model

$\lambda(t_j,\boldsymbol{z}_i)$ is the hazard at time t_j for an individual i with covariate values $\boldsymbol{z}_i$ and $\lambda_0(t_j)$ is the baseline hazard at time t_j. Notice that $\lambda_0(t)$ is the discrete baseline hazard function and can be understood as the hazard function for the set of covariates $\boldsymbol{z} = 0$, or without incorporating predictors in equation (2.33), once:

$$\lambda(t,\boldsymbol{z}=0) = y^{-1}[y[\lambda_0(t)]] = \lambda_0(t) \tag{2.34}$$

The hazard ratio (HR) is defined as the rate between the hazard function of two individuals a y b with covariate vectors $\boldsymbol{z}_a$ and $\boldsymbol{z}_b$ and has the general shape:

$$HR(t_i,\boldsymbol{z}_a,\boldsymbol{z}_b) = \frac{\lambda(t_j,\boldsymbol{z}_a)}{\lambda(t_j,\boldsymbol{z}_b)} = \frac{y^{-1}[y[\lambda_0(t_j)] + \boldsymbol{z}_a{}^t\boldsymbol{\beta}]}{y^{-1}[y[\lambda_0(t_j)] + \boldsymbol{z}_b{}^t\boldsymbol{\beta}]} \tag{2.35}$$

The interpretation of this result is that individuals with features $\boldsymbol{z}_a$ have a risk $HR(t_i,\boldsymbol{z}_a,\boldsymbol{z}_b)$ times greater/less than individuals with features $\boldsymbol{z}_b$. If $HR(t_i,\boldsymbol{z}_a,\boldsymbol{z}_b) < 1$ then features $\boldsymbol{z}_a$ retards the failure time with respect to $\boldsymbol{z}_b$, if $HR(t_i,\boldsymbol{z}_a,\boldsymbol{z}_b) = 1$ or is very close to 1 then the effect over the failure time is the same, if $HR(t_i,\boldsymbol{z}_a,\boldsymbol{z}_b) > 1$ then features $\boldsymbol{z}_a$ accelerates the failure time with respect to $\boldsymbol{z}_b$.

The Cox proportional-hazards model or standard discrete relative risk model comes from equation (2.32) by using the logarithm function:

$$log(\lambda(t_j,\boldsymbol{z}_i)) = log(\lambda_0(t_j)) + \boldsymbol{z}_i{}^t\boldsymbol{\beta} \tag{2.36}$$

Taking exponential, this model defines the hazard as

$$\lambda(t_j,\boldsymbol{z}_i) = \lambda_0(t_j)e^{\boldsymbol{z}_i{}^t\boldsymbol{\beta}}. \tag{2.37}$$

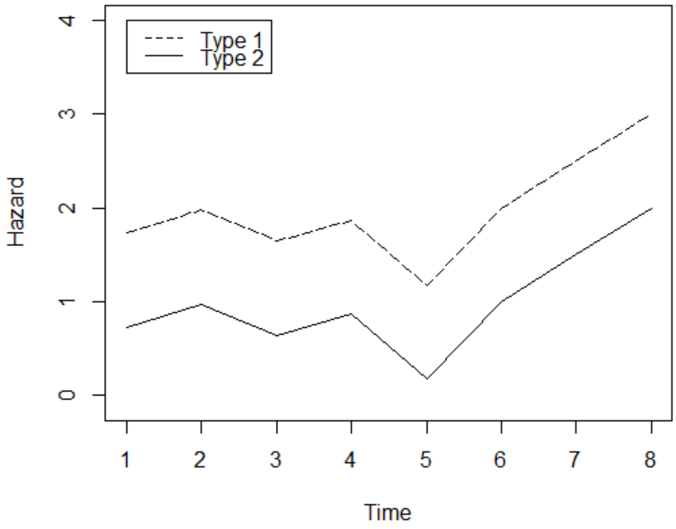
The model has two parts: a parametric one, depending on the vector of covariates or factors $e^{\boldsymbol{z}_i{}^t\boldsymbol{\beta}}$; and one nonparametric part, depending on the time, $\lambda_0(t)$, but without specifying a shape for the distribution of the survival times. That is why this model is described as a semiparametric model or partially parametric (Allison (1982)) and is just this feature which makes it robust, making this model a good alternative under the lack of knowledge about the real distribution of the failure times in analysis. Another important feature is that the linear predictor is associated with the hazard function through the exponential function that guarantees nonnegative values. Note that the covariates can be independent or time dependent.

In this model the HR for two individuals a y b, with covariate vectors $\boldsymbol{z}_a$ and $\boldsymbol{z}_b$, comes from equation (2.35) by using the logarithm function and has the shape:

$$HR(t, \boldsymbol{z}_a, \boldsymbol{z}_b) = \frac{\lambda_0(t)e^{\boldsymbol{z}_a{}^t\beta}}{\lambda_0(t)e^{\boldsymbol{z}_b{}^t\beta}} = \frac{e^{\boldsymbol{z}_a{}^t\beta}}{e^{\boldsymbol{z}_b{}^t\beta}} = e^{(\boldsymbol{z}_a-\boldsymbol{z}_b)^t\beta} \tag{2.38}$$

Notice that this result does not depend on the baseline hazard function, but only on the covariates vector and the regression coefficients. The interpretation of this results is explained with more details in the appendix (A.0.2). When the Cox model is fitted it is necessary to verify that the proportionality assumption holds. Graphical methods can be used to assure that the effect of a variable is constant over time, for example, comparing hazard curves of individuals with different variables and confirming that they are proportional (see Figure (**??**)). Appendix (A.0.3) shows an example of the lack of this assumption. Schoenfeld test of residuals (Schoenfeld (1980)) is another method that can be used to test the proportional hazard assumption.

Figure 2.3: Proportionality hazard assumption in the Cox model: In this example two hazard functions (for two types of individuals) are shown, they are proportional along time.



In case of a bivariate analysis for a covariate z_j the HR represents the rate of hazards when the variable increases in one unit. Then, if we need to analyze the HR when the continuous variable increases in c units the formula (A.5) will consider the values cz_a and cz_b, instead of

z_a and z_b, respectively.

The confidence interval for the hazard ratio is given by

$$e^{\hat{\beta}\pm z_{\alpha/2}\sqrt{Var(\hat{\beta})}} = e^{\hat{\beta}\pm z_{\alpha/2}se(\hat{\beta})} \tag{2.39}$$

where $e^{\hat{\beta}}$ is the effect of a covariate (without interactions).

If we increase a covariate z_i in c units the hazard ratio will be $HR = e^{(c\hat{\beta}_i)}$ and its associated confidence interval will be

$$e^{c\hat{\beta}_i \pm z_{\alpha/2}|c|se(\hat{\beta}_i)} \tag{2.40}$$

The discrete logistic model comes from equation (2.32) by using the logistic transformation (Logit or log odds):

$$log\left(\frac{\lambda(t_j, z_i)}{1-\lambda(t_j, z_i)}\right) = log\left(\frac{\lambda_0(t_j)}{1-\lambda_0(t_j)}\right) + {z_i}^t\beta \tag{2.41}$$

Taking exponential and solving

$$\frac{\lambda(t_j, z_i)}{1-\lambda(t_j, z_i)} = \frac{\lambda_0(t_j)}{1-\lambda_0(t_j)}e^{{z_i}^t\beta}. \tag{2.42}$$

This model defines the hazard as

$$\lambda(t_j, z_i) = \frac{\lambda_0(t_j)e^{{z_i}^t\beta}}{\lambda_0(t_j)e^{{z_i}^t\beta}+1-\lambda_0(t_j)}. \tag{2.43}$$

The main benefit of using a Logit transformation is its interpretation based on the odd of failure. The idea behind the odds is to measure how more likely is success than failure, which in this case means to measure the rate between the probability that the event occurs at time j given that did not occur before (λ_j) divided by the probability that the event does not occur at j given that did not occur before $(1-\lambda_j)$:

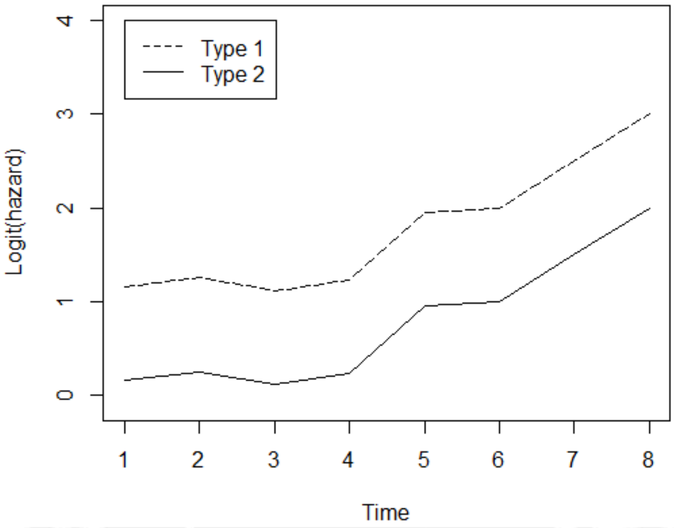
$$odds = \frac{\lambda_j}{1-\lambda_j} \tag{2.44}$$

Odds metric however is asymmetric and hence can not provide a fair curve to model the changing of probability of failure, to be used in the regression. The solution to deal with this scenario is by taking the natural logarithm of the odds, value that depends on the magnitude of the hazard.

$$logit(\lambda_j) = log(odds(\lambda_j)) = log\left(\frac{\lambda_j}{1-\lambda_j}\right) \tag{2.45}$$

In Logit scale if two values of hazard are small or close to 0 this transformation increase the difference between them and if they are large or far from 0 the difference will be decreased. In the Logit scale the size of the gap between two hazard functions is stable over time, being this condition known as proportionality assumption (see Figure (2.4)).

Figure 2.4: Proportionality Logit assumption in the logistic model. In this example two Logit(hazard) functions (for two types of individuals) are shown, they are proportional along time.



Some of the characteristics of this model are: (i) for each set of values of the covariates we can estimate a hazard function; (ii) the generated logit hazard functions based in the same set of covariates are parallel, then each function has the same shape for different set of values, although we can specify the shape of the hazard function that can be flat, linear or to has another general form; and (iii) the character of the covariates or predictors can be: time-invariant (regarding variables that keep fixes values over time) or time-varying (regarding variables which values can change over time).

The hazard function to be estimated will have two parts: a baseline Logit hazard function, and the effect of the predictors as a constant value to be summarized to the value of the baseline in each period. To provide flexibility to the shape of the model to be specified, instead of constraining the Logit hazard function to be linear with time (for instance having the component $\beta_0 + \beta_1 z_{time}$ in the model), a set of time indicators $\boldsymbol{D}$ is used to represent the dichotomized covariate of failure time, each one of them indexes one record of the observation-period data set, pointing out whether or not the individual took part of the observation at each time period (See table (2.1)).

The logistic hazard model can be also written as

$$\begin{aligned} \text{logit } \lambda(t_j, \boldsymbol{z}) &= \boldsymbol{D}^t\alpha + \boldsymbol{z}^t\beta \\ &= [\alpha_1 D_1 + \alpha_2 D_2 + ... + \alpha_j D_j] + [\beta_1 z_1 + \beta_2 z_2 + ... + \beta_j z_j] \qquad (2.46) \end{aligned}$$

Table 2.1: Mapping the failure time as time indicators

Period (from 1 to failure time)	D_1	D_2	...	D_{j-1}	D_j
1	1	0	0	0	0
2	0	1	0	0	0
...	0	0	...	0	0
$j-1$	0	0	0	1	0
j	0	0	0	0	1

*Each indicator shows “1” only at the time it is indexing

and rewriting in terms of the hazard function, 2.43 reduces to

$$\begin{aligned}\lambda(t_j, \boldsymbol{z}) &= \frac{1}{1+e^{-(\boldsymbol{D}^t\alpha+\boldsymbol{z}^t\beta)}} \\ &= \frac{1}{1+e^{-([\alpha_1D_1+\alpha_2D_2+...+\alpha_jD_j]+[\beta_1z_1+\beta_2z_2+...+\beta_jz_j])}} \qquad (2.47)\end{aligned}$$

The component $(\boldsymbol{D}^t\alpha)$ represents the baseline Logit hazard function. It will provide a set of values that can be understood as multiple intercept parameters (one per period) and can be interpreted as the hazard function without incorporating predictors (the set of P covariates $\boldsymbol{z} = 0$). Each one of them $\alpha_1, \alpha_2, ..., \alpha_j$ is the log odds of the hazard value in the respective period and we consider them as intercepts because for each individual i, when $D_{ix} = 1$ $(x = 1 \dots j)$ all other terms become 0 and the log odds of the hazard value for the xth period becomes α_x. Each slope parameter β measures the change in the log odds of failure (Logit values) per unit of difference in its associated covariate, assuming that the other covariates are constant.

For instance, considering as a dichotomy covariate z_{x1}, then we have the following models for $z_{x1} = 0$ and $z_{x1} = 1$, respectively:

$$\begin{aligned}\text{logit } \lambda(t_j, \boldsymbol{z}) &= [\alpha_1D_1+\alpha_2D_2+...+\alpha_jD_j] \\ \text{logit } \lambda(t_j, \boldsymbol{z}) &= [\alpha_1D_1+\alpha_2D_2+...+\alpha_jD_j]+\beta_{x1} \qquad (2.48)\end{aligned}$$

then, the respective values of Logit hazard for in the period j for these models are:

$$\begin{aligned}\text{logit } \lambda(t_j, \boldsymbol{z}) &= \alpha_j \\ \text{logit } \lambda(t_j, \boldsymbol{z}) &= \alpha_j+\beta_{x1} \qquad (2.49)\end{aligned}$$

Considering as a continuous covariate z_{x2}, in addition to z_{x1}, we have a model that incorporates more than one predictor and from different types, as follows:

$$\text{logit } \lambda(t_j, \boldsymbol{z}) = [\alpha_1D_1+\alpha_2D_2+...+\alpha_jD_j]+\beta_{x1}z_{x1}+\beta_{x2}z_{x2} \qquad (2.50)$$

In this case the baseline hazard function is given by the model when $z_{x1} = 0$ and $z_{x2} = 0$. The interpretation of the slope parameters β_{x1} and β_{x2} depend of the group of covariates included in the model. β_{x1} is the change in the log odds of failure, per unit of difference in the predictor z_{x1}, when z_{x2} remains constant.

The grouped relative risk model [2]comes from equation (2.32) by using the complementary log-log transformation (cloglog):

$$log(-log(1-\lambda(t_j,z_i))) = log(-log(1-\lambda_0(t_j))) + {z_i}^t\beta \tag{2.51}$$

Taking exponential twice we have

$$\begin{aligned} -log(1-\lambda(t_j,z_i)) &= -log(1-\lambda_0(t_j))e^{{z_i}^t\beta} \\ 1-\lambda(t_j,z_i) &= [e^{log(1-\lambda_0(t_j))}]^{e^{{z_i}^t\beta}} \end{aligned} \tag{2.52}$$

Isolating $\lambda(t_i,z_i)$, the grouped relative risk model defines the hazard as

$$\lambda(t_j,z_i) = 1-[1-\lambda_0(t_j)]^{e^{{z_i}^t\beta}}. \tag{2.53}$$

2.2.4 Estimation with covariates

We define in general terms the estimation process to be applied to the showed models. Let $\gamma_j = y[\lambda_0(t_j)]$, where $j = 1,...,k$, $\lambda_{k+1} = 1$ (all observations fail at the end of the study), and $g = y^{-1}$, the model can be written as

$$\lambda(t_j,z) = g[\gamma_j + z^t\beta] \tag{2.54}$$

The survival function in terms of the relative risk, can be written as

$$\hat{S}_j = \prod_{i=1}^{j}(1-g[\hat{\gamma_i} + z^t\hat{\beta}]). \tag{2.55}$$

Note that if $\hat{\beta} = 0$ it reduces to the Kaplan-Meier estimator (See equation (2.5)).

The likelihood function is built by multiplying the contributions for k failure times, as follows

$$L(\lambda \mid \gamma,\beta) = \prod_{i=1}^{k} L_i(\boldsymbol{\gamma},\boldsymbol{\beta}) \tag{2.56}$$

Assuming independent censoring the likelihood of $\boldsymbol{\gamma} = (\gamma_1,\gamma_2,...,\gamma_k)$, $\boldsymbol{\beta} = (\beta_1,\beta_2,...,\beta_p)$,

[2] Further explanation about the grouped relative risk model is out of the scope of the thesis

and considering l as the index for periods, can be written as

$$\begin{aligned} L(\lambda \mid \gamma, \beta) &= \prod_{i=1}^{k}\left[\prod_{l\in D_i}\lambda(t_i, z_l)\times\prod_{l\in R_i}(1-\lambda(t_i, z_l))\right] \\ &= \prod_{i=1}^{k}\left[\prod_{l\in D_i} g[\gamma_i + {z_l}^t\beta]\times\prod_{l\in R_i}(1-g[\gamma_i + {z_l}^t\beta])\right] \qquad (2.57) \\ & \qquad (2.58) \end{aligned}$$

and the likelihood function as

$$l(\lambda \mid \gamma, \beta) = \sum_{i=1}^{k}\left[\sum_{l\in D_i}\log(g[\gamma_i + {z_l}^t\beta]) + \sum_{l\in R_i}\log(1-g[\gamma_i + {z_l}^t\beta])\right] \qquad (2.59)$$

where D_i represents the indexes of observations that fail and R_i the indexes of censored observations at period i. The score function is defined as

$$S(\gamma, \beta) = \frac{\partial log L(\lambda \mid \gamma, \beta)}{\partial \beta_j} = \frac{\partial l(\lambda \mid \gamma, \beta)}{\partial \beta_j} \qquad (2.60)$$

where $j = 1, ..., p$. Then, the MLE are found by solving

$$S(\gamma, \beta) = \left(\frac{\partial l}{\partial \gamma_1}, ..., \frac{\partial l}{\partial \gamma_k}, \frac{\partial l}{\partial \beta_1}, ..., \frac{\partial l}{\partial \beta_p}\right) = 0 \qquad (2.61)$$

where each component has the shape

$$\frac{\partial l}{\partial \gamma_i} = \sum_{l\in D_i}\frac{g'_{il}}{g_{il}} - \sum_{l\in R_i}\frac{g'_{il}}{1-g_{il}} \qquad (2.62)$$

and

$$\frac{\partial l}{\partial \beta_u} = \sum_{i=1}^{k}\left[\sum_{l\in D_i}\frac{z_{lu}(g'_{il})}{g_{il}} + \sum_{l\in R_i}\frac{z_{lu}(g'_{il})}{1-g_{il}}\right] \qquad (2.63)$$

being $g_{il} = g[\gamma_i + {z_i}^t\beta]$, $1 \le i \le k$ and $1 \le u \le p$.

A Newton-Raphson iteration algorithm can be included to update the current values (γ_0, β_0) to (γ_1, β_1) until convergence according to the following relation

$$(\gamma_1, \beta_1) = (\gamma_0, \beta_0) + H_0^{-1}S(\gamma_0, \beta_0) \qquad (2.64)$$

where H_0 represent H at the point (γ_0, β_0). After this numerical methods we get the MLE estimators $\hat{\boldsymbol{\gamma}} = (\hat{\gamma}_1, ..., \hat{\gamma}_k)$ and $\hat{\beta} = (\hat{\beta}_1, ..., \hat{\beta}_p)$.

To obtain the variance of the estimators, the observed Fisher information is calculated. For this we need the negative of the second derivative (Hessian matrix) of the log-likelihood function evaluated at the MLE of $\boldsymbol{\gamma}$ and $\boldsymbol{\beta}$, i.e.,

$$H = I(\hat{\boldsymbol{\gamma}}, \hat{\boldsymbol{\beta}}) = \frac{\partial^2 logL(\lambda \mid \hat{\gamma}, \hat{\beta})}{\partial\gamma\partial\beta_j} = -\frac{\partial^2 l(\lambda \mid \hat{\gamma}, \hat{\beta})}{\partial\gamma\partial\beta_j} \tag{2.65}$$

where

$$H = \begin{pmatrix} H_{11} & H_{12} \\ H_{21} & H_{22} \end{pmatrix} = \begin{pmatrix} \frac{-\partial^2 l}{\partial\gamma^2} & \frac{-\partial^2 l}{\partial\gamma\partial\beta} \\ \frac{-\partial^2 l}{\partial\gamma\partial\beta} & \frac{-\partial^2 l}{\partial\beta^2} \end{pmatrix} \tag{2.66}$$

An estimator of the variance-covariance matrix of $(\hat{\boldsymbol{\gamma}}, \hat{\boldsymbol{\beta}})$, $\hat{\sum}$, is given by $I(\hat{\boldsymbol{\gamma}}, \hat{\boldsymbol{\beta}})^{-1}$.

2.2.5 Particulars for Cox regression

In the Cox regression model the regression parameter vector $\boldsymbol{\beta} = (\beta_1, \beta_2, ..., \beta_p)$ is estimated using the partial likelihood function (PL), which is based on the product of the likelihoods of all failures occurred, unlike the ordinary maximum likelihood estimation that is based on the product of likelihoods for all individuals in the sample. In the estimation only take part the probabilities of real failure times but not those of censored data, although at calculating the probabilities of failure times it takes into account all observations under risk of failure, censored or not in the future (Cox (1975)). By this procedure we reach to complete the parametric component of the model, being enough to do inference over that vector of parameters and calculating the hazard ratios, despite of not being a real likelihood function.

Let L or $L(\beta_1, ..., \beta_p)$ be the partial likelihood function, n the number of observations, k the number of failure times (there are no ties), $n-k$ the number of censored times, $t_1, ..., t_k$ the ordered failure times, R_i (for $i = 1, ..., k$) the number of observations under risk at t_i, and L_i the portion of the likelihood function because of the contribution of t_i (for $i = 1, ..., k$). Then the definition of L_i become as follows:

$$\begin{aligned} L_i(\boldsymbol{\beta}) = \frac{\lambda(t_i, z_i)}{\sum_{l \in R_i} \lambda(t_i, z_l)} &= \frac{\lambda_0(t_i) e^{z_i^t \boldsymbol{\beta}}}{\sum_{l \in R_i} \lambda_0(t_i) e^{z_l^t \boldsymbol{\beta}}} \\ &= \frac{e^{z_i^t \boldsymbol{\beta}}}{\sum_{l \in R_i} e^{z_l^t \boldsymbol{\beta}}} \end{aligned} \tag{2.67}$$

After obtaining the contribution to the likelihood by each failure time we have the following expression to be used as a likelihood function:

$$L(\boldsymbol{\beta}) = \prod_{i=1}^{k} L_i(\boldsymbol{\beta}) = \prod_{i=1}^{k} \left[\frac{e^{\mathbf{z}_i^t \beta}}{\sum_{l \in R_i} e^{\mathbf{z}_l^t \boldsymbol{\beta}}} \right] \tag{2.68}$$

In a real situation, either due to the lack of accurate instruments to measure continuous data or because of the data seems to be discrete, we could observe more than one failure at the same time. In this way, there have been some **approximations of the partial likelihood function in case of ties**. These approaches aim to estimate parameters of the hazard function making some adaption to the likelihood function defined above. Let m_i be the multiplicity or number of observations that fail at t_i, then $m_i = 1$ if only one observation fail and $m_i > 1$ if instead more than one fail. If we denote u_j to the random selection of m_i observations that fail having R_i elements at risk, then the number of possible subsets u_j's is given by:

$$\binom{R_i}{m_i} = \frac{R_i!}{m_i!(R_i m_i)!} \tag{2.69}$$

Let $\boldsymbol{U_i}$ be the set of all possible subsets where $\boldsymbol{U_i} = (u_1, ..., u_{\binom{R_i}{m_i}})$; $\boldsymbol{z_k} = (z_{1k}, z_{2k}, ..., z_{pk})$ be the covariates vector of the kth observation, $\boldsymbol{x_{u_j}} = \sum_{k \in u_j} \boldsymbol{z_k} = (x_{1u_j}, x_{2u_j}, ..., x_{pu_j})$ be the vector of summarization of covariates where x_{lu_j} is the summarization of the lth covariate for the m_i observations within u_j; u_i^* be the subset of m_i observations that fail at t_i; and $\boldsymbol{x_{u_i^*}} = \sum_{k \in u_i^*} \boldsymbol{z_k} = (x^*_{1u_i^*}, x^*_{2u_i^*}, ..., x^*_{pu_i^*})$, where $x_{lu_j^*}$ is the summarization of the lth covariate of the m_i observations within u_i^* that fail at t_i. Then, when m_i is a small number in comparison with r_i we can use the following two approaches to the likelihood functions.

The Breslow (1974) approach defines the likelihood function as:

$$L_B(\boldsymbol{\beta}) = \prod_{i=1}^{k} \left[\frac{e^{\mathbf{x}^t_{u_i^*}\beta}}{\left[\sum_{l \in R_i} e^{\mathbf{z}_l^t \boldsymbol{\beta}}\right]^{m_i}} \right] \tag{2.70}$$

The Efron (1977) approach defines the likelihood function as:

$$L_E(\boldsymbol{\beta}) = \prod_{i=1}^{k} \left[\frac{e^{\mathbf{x}^t_{u_i^*}\beta}}{\prod_{j=1}^{m_i} \left[\sum_{l \in R_i} e^{\mathbf{z}_l^t \boldsymbol{\beta}} - [\frac{j-1}{m_i}] \sum_{l \in u_i^*} e^{\mathbf{z}_l^t \boldsymbol{\beta}}\right]} \right] \tag{2.71}$$

2.2.6 Particulars for logistic regression

The method gets the estimated values of $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ that maximize the likelihood of the observed sample data. The likelihood function models the probability of getting the sample currently observed. In order to fit this function it is necessary to format the data set in an "individual-period" level which means that we will have the following scheme of covariates: PERIOD, representing the number of period; EVENT (E), representing whether or not a failure has occurred at each period; and a series of time indicators $\boldsymbol{D}$ (See table (2.1)). For example, table 2.3 shows how data can be rewritten at the commented level in comparison with data in 2.2, where n is the number of observations with failure times $t_1, ..., t_n$ and censoring indicators $c_1, ..., c_n$, k is the number of failure times and $\boldsymbol{z}$ represents a vector of fixed predictors.

The likelihood function for the discrete-time hazard aims to model the specific probability of observing a specific pattern of 0s and 1s for the variable EVENT in an "Individual-period level" data set, and the output of the estimation process is a set of values for vectors of coefficients αs and βs. Let J_i the number of observations for individual i (one per period

Table 2.2: "Individual level" example data set

ID	T	Censor	z_b	z_f	z_{v1}	z_{v2}	z_{v3}	...	z_{vt_2}	...	z_{vt_n}	...	z_{vt_1-1}	z_{vt_1}
1	5	1	1	z_{f1}	z_{v1_1}	z_{v2_1}	z_{v3_1}	...	$z_{vt_{2}1}$	...	z_{vt_n1}	...	$z_{vt_1-1_1}$	z_{vt_11}
2	3	0	0	z_{f2}	1	0	0	...	0	-	-	-	-	-
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
n	7	1	1	z_{fn}	z_{v1_n}	z_{v2_n}	z_{v3_n}	...	$z_{vt_{2n}}$	...	$z_{vt_{nn}}$	-	-	-

Table 2.3: "Individual-period level" example data set

ID	Period	Event	D_1	D_2	D_3	D_4	D_5	D_6	D_7	...	D_{t_1-1}	D_{t_1}	z_b	z_f	z_v
1	1	0	1	0	0	...	0	...	0	...	0	0	1	z_{f1}	z_{v1_1}
1	2	0	0	1	0	...	0	...	0	...	0	0	1	z_{f1}	z_{v1_2}
1	3	0	0	0	1	...	0	...	0	...	0	0	1	z_{f1}	z_{v1_3}
1	4	0	0	0	0	...	1	...	0	...	0	0	1	z_{f1}	$z_{v1_{t_2}}$
1	5	1	0	0	0	...	0	...	1	...	0	0	1	z_{f1}	$z_{v1_{t_n}}$
2	1	0	1	0	0	...	0	...	0	...	0	0	0	z_{f2}	1
2	2	0	0	1	0	...	0	...	0	...	0	0	0	z_{f2}	0
2	3	0	0	0	1	...	0	...	0	...	0	0	0	z_{f2}	0
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
n	1	0	1	0	0	...	0	...	0	...	0	0	1	z_{fn}	z_{vn_1}
n	2	0	0	1	0	...	0	...	0	...	0	0	1	z_{fn}	z_{vn_2}
n	3	0	0	0	1	...	0	...	0	...	0	0	1	z_{fn}	z_{vn_3}
n	4	0	0	0	0	1	0	...	0	...	0	0	1	z_{fn}	z_{vn_3}
n	5	0	0	0	0	...	1	...	0	...	0	0	1	z_{fn}	$z_{vn_{t_2}}$
n	6	0	0	0	0	...	1	...	0	...	0	0	1	z_{fn}	$z_{vn_{t_2}}$
n	7	1	0	0	0	...	0	...	1	...	0	0	1	z_{fn}	$z_{vn_{t_n}}$

* In the example z_b represents a binary variable, z_f is fixed value and z_v a time-varying variable

of risk), then the number of terms of the likelihood function will be equal to the number of rows in the dataset and we will assume that the J_i observations for each individual are independent. The contribution to the likelihood function of an individual i who fails at time t_j is $\lambda(t_{ij}, \boldsymbol{z})$, and the contribution of an individual that does not fail at time t_j is $(1-\lambda(t_{ij}, \boldsymbol{z}))$. Then is easy to see that individuals that fail contribute with one term of the first type and with $J_i - 1$ terms of the second type, while censored observations contribute with J terms of the second type. Using this explanation the likelihood function L or $L(\alpha_1, ..., \alpha, \beta_1, ..., \beta_p)$ has the form

$$\begin{aligned}
L(\underline{\lambda} \mid t_1, t_2, .., t_n) &= \prod_{i=1}^{n}\prod_{j=1}^{J_i}\Big[\lambda(t_j, \boldsymbol{z})\Big]^{E_{ij}}\Big[1-\lambda(t_j, \boldsymbol{z})\Big]^{(1-E_{ij})} \\
&= \prod_{i=1}^{n}\prod_{j=1}^{J_i}\left[\frac{1}{1+e^{-(\boldsymbol{D}^t\alpha+\boldsymbol{z}^t\beta)}}\right]^{E_{ij}}\left[1-\frac{1}{1+e^{-(\boldsymbol{D}^t\alpha+\boldsymbol{z}^t\beta)}}\right]^{(1-E_{ij})} \\
&= \prod_{i=1}^{n}\prod_{j=1}^{J_i}\left[\frac{1}{1+e^{-([\alpha_1 D_1+\alpha_2 D_2+...+\alpha_j D_j]+[\beta_1 z_1+\beta_2 z_2+...+\beta_j z_j])}}\right]^{E_{ij}} \times \\
&\quad \left[1-\frac{1}{1+e^{-([\alpha_1 D_1+\alpha_2 D_2+...+\alpha_j D_j]+[\beta_1 z_1+\beta_2 z_2+...+\beta_j z_j])}}\right]^{(1-E_{ij})} \qquad (2.72)
\end{aligned}$$

Then the log likelihood function can be written as

$$l(\underline{\lambda} \mid t_1, t_2, .., t_n) = \sum_{i=1}^{n} \sum_{j=1}^{J_i} E_{ij} log \lambda(t_j, \boldsymbol{z}) + (1 - E_{ij}) log(1 - \lambda(t_j, \boldsymbol{z})) \tag{2.73}$$

2.2.7 Hypothesis testing

The asymptotic distribution of the MLE estimator $\hat{\beta} = (\hat{\beta}_1, ..., \hat{\beta}_p)$ is normal with mean β and covariance matrix $\hat{\sum} = Var(\hat{\beta}) = I(\hat{\beta})^{-1}$, which provides the conditions to apply tests from generalized linear models.

To test the hypothesis $H_0 : \beta_j = 0$ versus $H_1 : \beta_j \neq 0$ we can use the Wald statistic, defined as the rate between the estimator and its standard deviation:

$$z_W = \frac{\hat{\beta}_j}{\sqrt{Var(\hat{\beta}_j)}} \tag{2.74}$$

Under the null hypothesis one can show that z_W follows a standard normal distribution and, therefore the $100(1 - \alpha)$ % asymptotic confidence interval for the estimator $\hat{\beta}_j$ is given by

$$\hat{\beta}_j \pm z_{1-\alpha/2} \sqrt{Var[\hat{\beta}_j]} \tag{2.75}$$

To test the hypothesis $H_0 : \boldsymbol{\beta} = \boldsymbol{\beta_0}$ versus $H_1 : \boldsymbol{\beta} \neq \boldsymbol{\beta_0}$ one can alternative use the Wald test, the likelihood ratio test or the score or Logrank test. Under the null hypothesis we assume that there is no relation between the variable (in the univariate case), or one of the variables (in the multivariate case), and the survival time ($\beta = 0$). The statistic that provides the easiest interpretation is given by the Wald test which, as we will see, does not use the likelihood function. On the other hand, as it shows the Log Rank test only uses the coefficients under the null hypothesis, then if it is a goal to estimate many coefficients this method will be faster than the others. Finally the Likelihood ratio test converges faster to the normal distribution.

The Wald test use the fact that the distribution of the MLE estimator $\hat{\beta} = (\hat{\beta}_1, ..., \hat{\beta}_p)$ is normal with mean β and covariance matrix $\sum$, that can be estimated by $\hat{\sum} = Var(\hat{\beta}) = I(\hat{\beta})^{-1}$. Under that null hypothesis the Wald statistic follows a χ^2 distribution with p degrees of freedom and it is defined as

$$\chi_W = (\hat{\boldsymbol{\beta}} - \boldsymbol{\beta_0})^t I(\hat{\boldsymbol{\beta}})(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta_0}) \tag{2.76}$$

The Likelihood ratio test calculates an statistic that use the partial likelihood function evaluated in $\hat{\boldsymbol{\beta}}$, $L(\hat{\boldsymbol{\beta}})$, and evaluated in $\boldsymbol{\beta_0}$, $L(\boldsymbol{\beta_0})$. Under that null hypothesis the statistic follow a χ^2 distribution with p degrees of freedom and is defined as

$$\begin{aligned} \chi_{LR} &= 2(logL(\hat{\boldsymbol{\beta}}) - logL(\boldsymbol{\beta_0})) \\ &= 2(l(\hat{\boldsymbol{\beta}}) - l(\boldsymbol{\beta_0})) \end{aligned} \tag{2.77}$$

where l is the log likelihood function.

The Score or Log Rank test use the gradient of the logarithm of the partial likelihood function evaluated at the null hypothesis ($\boldsymbol{\beta} = \boldsymbol{\beta_0}$). Under the condition that the scores vector follows approximately a normal distribution with means 0 and covariance matrix $I(\boldsymbol{\beta})$. Furthermore, the statistic follows a χ^2 distribution with p degrees of freedom and it is defined by

$$\chi_S = \left(\frac{\partial L(\boldsymbol{\beta_0})}{\partial \boldsymbol{\beta}}\right)^t \left(\frac{\partial L^2(\boldsymbol{\beta_0})}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}^t}\right)^{-1} \frac{\partial L(\boldsymbol{\beta_0})}{\partial \boldsymbol{\beta}} \tag{2.78}$$

2.2.8 Comparison of models with the likelihood test

Let us suppose we estimated two models, one with p predictors and a second one with $p+q$ predictors. The vector of parameters for the first one is $\hat{\beta} = (\hat{\beta}_1, ..., \hat{\beta}_p)$ and for the second one is $\hat{\beta} = (\hat{\beta}_1, ..., \hat{\beta}_{p+q})$. Two methods are reviewed to compare the goodness of fit over a group of models.

- The Deviance statistic. In the first model we have the hypothesis $H_0 : \beta_j = 0$ versus $H_1 : \beta_j \neq 0$ (at least one coefficient different to zero), where $j = 1, ..., p$; and for the second model we have the contrast $H_0 : \beta_j = 0$ versus $H_1 : \beta_j \neq 0$, where $j = 1, ..., p+q$. Then we can calculate uses the deviance concept $D_i = -2logL_i(\hat{\boldsymbol{\beta}}) = -2l_i(\hat{\boldsymbol{\beta}})$. The statistic follows a χ^2 distribution with q degrees of freedom and is defined as

$$\chi_{DD} = D_1 - D_2 = D = -2logL_1(\hat{\boldsymbol{\beta}}) - -2logL_2(\hat{\boldsymbol{\beta}}) = -2log\left(\frac{L_1(\hat{\boldsymbol{\beta}})}{L_2(\hat{\boldsymbol{\beta}})}\right) \tag{2.79}$$

- The Akaike's information criterion (Akaike (1973)). It is a metric for comparing models which measures both, how well the model fits the data, and how complex the model is. It uses the parsimony principle that says that a model using fewer parameters and explaining the context almost in a same level is the best. This statistic is defined as

$$\begin{aligned} AIC(\hat{\boldsymbol{\beta}}) &= -2logL(\hat{\boldsymbol{\beta}}) + 2p \\ &= -2l(\hat{\boldsymbol{\beta}}) + 2p. \end{aligned} \tag{2.80}$$

  Then, AIC_i values of R models ($i = 1, \ldots, R$) are compared by calculating the differences between each AIC_i and the minimum AIC value $AIC_{min} = argmin(AIC_i)$: $\Delta_i = AIC_i - AIC_{min}$. The maximum difference will point out which model provides the most plausibility according with the following scale: If $\Delta_i \in (0-2)$ then there is a similar plausibility, else if $\Delta_i \in (4-7)$ then there is less plausibility, and if $\Delta_i \geq 10$ then there is much less plausibility. The AIC metric is recommended under the condition $n/p \geq 40$, where n is the number of observations within the sample on study. If the condition is not reached there is a variation of the metric named “Akaike second order information criterion”, AIC_c, defined as

$$AIC_c(\hat{\boldsymbol{\beta}}) = AIC(\hat{\boldsymbol{\beta}}) + \frac{2p(p+1)}{n-p-1.} \tag{2.81}$$

Chapter 3

Application

> *"There is an orderliness in the universe,*
> *there is an unalterable law governing everything*
> *and every being that exists or lives.*
> *It is no blind law, for no blind law can govern the conduct of living beings".*
> *Gandhi, 1928*

3.1 Framework of the study

The application in this work has been conducted at a private University in Perú. This consist of a population of 26,790 students enrolled between 2004 and 2012 in the undergraduate programs at the University (18 cohorts of incoming students in 9 years)[1]. After admission we classified a student to be enrolled in one of the following units: Arts, Architecture and Urbanism, Humanities and Social Sciences, Formal Sciences and Engineering, or Education. Dropout was defined as not being enrolled for at least two semesters (one year), excluding the summer academical cycles. Censored students in this context are due to graduation or due to they did not accomplish the dropout condition, within the observation period. Our primary objective is to study factors associated with time to drop out.

3.2 Data Structure

The covariates considered in the study were obtained at two levels that we will call the baseline level and the follow-up level.

In the first level we considered variables whose values do not change over time, also named fixed variables, as: gender (female, male), area of first enrollment (Arts, Architecture and Urbanism, Humanities and Social Sciences, Formal Sciences and Engineering, or Education), high school type (private, public, parochial, others ("fe y alegría",armed forces)), high school

[1] Data provided by the Department of Statistics and Institutional Intelligence (PUCP)

location (Lima y Callao, provinces, abroad), type of residence (own, renting, living with relatives, others (use, surveillance, rural or precarious occupation)), marital status (single, married and others (partners, religious, divorced)), indicator of whether or not the mother and father are alive (yes, no), mother and father education levels (college education [postgraduate, university, incomplete university], technical education [complete or incomplete technical education], high school [complete or incomplete high school], primary school [complete or incomplete primary school]), number of relatives at the University , indicator of whether or not the student was working at admission time, and modality of admission (entrance examination, entrance examination in first attempt, pre-university center, special programs of admission (bachelor diplomas, exonerations, admission by "Bicentennial of the Republic of Perú" program, agreements with "Fe y alegría" schools, admission by "R.P. Jorge Dintilhac SS.CC." program, scholar excellency program)).

In the second level we considered variables measured at follow-up whose values change over time, also named time−varying variables, as: payment scale, indicator to have asked for reclassification, type of benefit (scholarship, discount, loans, others), number of visits to health service, age, standardized academic performance index or "CRAEST" (equivalent metric for GPA), and percentage of passed credits.

3.3 Results

Of the $26,790$ individuals, 20881 (78%) were persistent students or were finally graduated, while 5909 (22%) were dropouts. In general, there is a balanced amount of female (43.42%) and male (56.58%) enrolled students. The majority of them was single (at least 94.62%). Ninety two percent of students was enrolled in Formal Sciences and Engineering, and Humanities and Social Sciences; 83% come from private or private religious schools; 85% studied in Lima and Callao; and 50.49% lives in their own houses (see table (3.1)). About family information, at least 91.49% of students has its mother alive and 81.05% its father alive, being the most common education level of parents the college education, 47.95% in the case of the mother and 59.43% in the case of the father. The 94.23% has other relatives (parents or siblings) studying at the University. Mostly of students, 44.24%, are admitted to the University by the entrance examination and only 8.98% is admitted by studying at the pre−university center (see table (3.2)). The 73.71% of students starts the university in the first semester of the year (see table (3.3)).

A general behavior of the dropout phenomena can be described by estimating the hazard and survival functions for the whole population of students using the Kaplan-Meier estimators (see Table (3.4)). Table (3.4) shows that since enrollment, 78.7% (95% CI: 0.782-0.793) of students remain until 6 semesters (3 years); 73.7% (95% CI: 0.731-0.744) of students remain until the 10th semester (5 years); 31.1% of students has dropped out the University after 7 years. In any period the hazard is not greater than 4.4% and this maximum scenario is reached in the 3rd semester. By the end of the data collection (second semester of 2012) 5909 students had left studying, and those 20881 who did not, were censored, representing

Table 3.1: Descriptive statistics at enrollment

	Population N = 26790	
Factor	N	Percentage (%)
Dropout		
Yes (event)	5909	22.00%
No (censored)	20881	78.00%
Gender		
Female	11633	43.42%
Male	15157	56.58%
Marital status		
Single	25348	94.62%
Married and others	26	<1.00%
Missing	1416	5.29%
Area of first enrollment		
Formal Sciences and Engineering	12028	44.90%
Architecture and Urbanism	1126	4.20%
Arts	779	2.91%
Education	292	1.09%
Humanities and Social Sciences	12565	46.90%
High school type		
Private / private religious	22143	83.00%
Public / public religious	2534	9.00%
Parochial	1294	5.00%
Others (“fe y alegria”, armed forces)	819	3.00%
High school location		
Lima y Callao	22771	85.00%
Provinces	3970	14.82%
Abroad	45	<1.00%
Missing	4	<1.00%
Type of housing		
Own	13525	50.49%
Renting	4485	16.74%
Living with relatives	6124	22.86%
Others (use, surveillance, rural or precarious occupation)	1183	4.42%
Missing	1473	5.50%

*academic semesters, **nuevos soles (S/.), $^+$interquartile range IQR [Q_1,Q_3]

Table 3.2: Descriptive statistics at enrollment

Factor	N	Percentage (%)
Indicator of mother alive		
Yes	24511	91.49%
No	835	3.12%
Missing	1444	5.39%
Indicator of father alive		
Yes	21713	81.05%
No	3633	13.56%
Missing	1444	5.39%
Mother education level		
College education (Postgraduate, university, incomplete university)	12846	47.95%
Technical education (Technical, incomplete technical)	5396	20.14%
High school (High school, incomplete high school)	6285	23.46%
Primary school (Primary school, incomplete primary school)	541	2.02%
Missing	1722	6.43%
Father education level		
College education (Postgraduate, university, incomplete university)	15922	59.43%
Technical education (Technical, incomplete technical)	3425	12.78%
High school (High school, incomplete high school)	4369	16.31%
Primary school (Primary school, incomplete primary school)	306	1.14%
Missing	2768	10.33%
Number of relatives at the University		
Only student	1546	5.77%
Two*	1943	7.25%
Three*	6985	26.07%
Four*	16281	60.77%
Five*	35	<1.00%
Working		
Yes	461	1.72%
No	15274	57.01%
Missing	11055	41.27%
Type of admission		
Entrance examination	11851	44.24%
Entrance examination − first attempt	6601	24.64%
Pre−university center	2407	8.98%
Special programs of admission	5931	22.14%

*values different to “Only student” are considering relatives and the student

Table 3.3: Descriptive statistics at enrollment

Factor	N	Percentage (%)
Cycle of entry		
First semester of the year	19746	73.71%
Second semester of the year	7044	26.29%
Cohort*		
2004 first semester of the year	1520	5.67%
2004 second semester of the year	842	3.14%
2005 first semester of the year	1672	6.24%
2005 second semester of the year	768	2.87%
2006 first semester of the year	1993	7.44%
2006 second semester of the year	668	2.49%
2007 first semester of the year	2560	9.56%
2007 second semester of the year	634	2.37%
2008 first semester of the year	2238	8.35%
2008 second semester of the year	712	2.66%
2009 first semester of the year	2500	9.33%
2009 second semester of the year	772	2.88%
2010 first semester of the year	2374	8.86%
2010 second semester of the year	977	3.65%
2011 first semester of the year	2239	8.36%
2011 second semester of the year	854	3.19%
2012 first semester of the year	2650	9.89%
2012 second semester of the year	817	3.05%

*Admission of students is not effective in summer academic cycles, only in semesters 1 and 2

the 78% of the analyzed population. In the appendix (B.0.5), Figure (B.1) shows graphically the explained description.

Results of a bivariate analysis (univariate regression) give a high level interpretation. We execute two estimations, using the Cox regression and the logistic regression obtaining similar coefficients. The objective of showing these two methods is to demonstrate that the estimations can be compared with fixed covariates and the analysis of time-varying covariates is only restricted to the logistic model. Taking as reference the logistic regression, outcomes show that in male students the odds of dropout hazard ("ODH") increases in 59% (OR = 1.59, 95% IC: 1.50-1.68) over female students. Taking as reference students enrolled in formal sciences and engineering, students enrolled in humanities and social sciences have 49% (OR = 0.51, 95% IC: 0.48-0.54) less ODH, while the enrolled ones in Arts has 22% (OR = 0.78, 95% IC: 0.67-0.92) less ODH. About the high school of origin, students who came from parochial schools have 23% (OR = 0.77, 95% IC: 0.67-0.89) less ODH than those who came

Table 3.4: Kaplan-Meier estimation for the entire population

Semester	At risk[a]	Failures[b]	Risk[c]	Survival[d]	95% CI
1st	26790	885	0.033	0.967	(0.965, 0.969)
2nd	24933	1076	0.043	0.925	(0.922, 0.928)
3rd	21264	937	0.044	0.884	(0.881, 0.888)
4th	19426	781	0.040	0.849	(0.844, 0.853)
5th	16656	642	0.039	0.816	(0.811, 0.821)
6th	15134	532	0.035	0.787	(0.782, 0.793)
7th	12722	322	0.025	0.768	(0.762, 0.773)
8th	11776	217	0.018	0.753	(0.748, 0.759)
9th	9712	100	0.010	0.746	(0.740, 0.752)
10th	9094	101	0.011	0.737	(0.731, 0.744)
11th	7040	78	0.011	0.729	(0.723, 0.736)
12th	5787	70	0.012	0.720	(0.714, 0.727)
13th	3385	48	0.014	0.710	(0.703, 0.717)
14th	2250	66	0.029	0.689	(0.681, 0.698)
15th	1133	33	0.029	0.669	(0.659, 0.680)
16th	644	21	0.033	0.647	(0.634, 0.661)
17th	267	0	0.000	0.647	(0.634, 0.661)
18th	103	0	0.000	0.647	(0.634, 0.661)

[a] number of students at risk of dropout

[b] number of dropouts in the current semester and before the next one

[c] probability to dropout in the current semester

[d] probability that a student remains at the University until the current semester

from a private or private religious schools. In the cases of “fe y alegría” schools and students from armed forces, that rate reduces to 22% (OR = 0.78, 95% IC: 0.65-0.93). Additionally, students who studied at the hight school in provinces have 13% (OR = 0.87, 95% IC: 0.80-0.95) less ODH than those who studied in Lima, while who studied abroad have 102% (OR = 2.02, 95% IC: 1.18-3.49) more ODH. About type of housing, we observed that in the case of students who rent the ODH increases in 17% (OR = 1.17, 95% IC: 1.08-1.26). On the other hand, for students who do not have any of their living parents the ODH increases, being the rate in the case of absent of the mother 36% (OR = 1.36, 95% IC: 1.17-1.58) and in the case of the father 10% (OR = 1.10, 95% IC: 1.02-1.20). The parent educational levels are inversely proportional to the ODH, while the level is higher the ODH decreases. The highest rates of ODH related with parents educational level is reached when they have studied only at the primary school, in the case of the mother the rate increases in 69% (OR = 1.69, 95% IC: 1.42-2.02) and in the case of the father increases in 61% (OR = 0.61, 95% IC: 1.27-2.03). Another outstanding feature is for the students who have some relatives associated with the University (current studies or in the past); in general for those students the ODH decreases.

If the student was working at the admission time the ODH increases in 45% (OR = 1.45, 95% IC: 1.17-1.80). In general, students who were admitted to the University by another modality different to the most common way, entrance examination, has less ODH. For instance, students belonging to the “top third percent” type of admission have 40% (OR = 0.60, 95% IC: 0.54-0.65) less risk to dropout. The relation between payment scale and ODH is directly proportional, if the level increases in one unit the ODH increases in 5% (OR = 1.05, 95% IC: 1.02-1.07) and students who asked for reclassification has less risk than those who did not in 22% (OR = 0.78, 95% IC: 0.67-0.92). In the same direction, students enjoying some benefit have less hazard. For example, in the case of loans, the rate decreases in 77% (OR = 0.23, 95% IC: 0.14-0.39). The age is also directly proportional to the ODH, each year a student increases its ODH in 15% (OR = 1.15, 95% IC: 1.13-1.17). The standardized academic performance index has an inverse effect, for each unit that it increases the ODH decreases in 15% (OR = 0.85, 95% IC: 0.84-0.85).

We point out now a criterion in order to select a group of variables in the multivariate model estimation. These variables have to be statistically significant at 25% . Given this rule we identify the following variables and their respective instances as significant ones: area of first enrollment (Humanities and Social Sciences, Education), high school type (public, parochial), type of residence (renting, living with relatives), indicator of whether or not the mother is alive, mother education level (technical education, high school, primary school), number of relatives at the University, modality of admission (entrance examination in first attempt, top third of the class in high school), payment scale, indicator of whether or not the student has asked for reclassification, number of visits to health service, standardized academic performance index, percentage of passed credits. In addition we are including the variable “age” ($p-$value $= 0.332$) as part of a research special feature, despite of not being include within the range of 25%. The decision was taken in order to get knowledge about how much more students increases their risk to drop out each year they are older.

After selecting a group of significant variables based in the univariate analysis and filtering those which in the multivariate model are not significant, we analyze those variables which are just in the limit of the significance level in order to confirm if we should or not keep them as part of a second instance or final multivariate model. Likelihood ratio tests are executed for the following variables, taking as reference a basic multivariate model that do not include any of them: high school type, indicator of whether or not the mother is alive, number of relatives at the University. Finally we decided to remove these three variables of the model after analyzing the obtained significance values for the hypothesis test. In appendix (A.0.4), table (??) shows the outcomes with which we conclude that there is no evidence to point out that two models are different in each test.

We keep the following variables as part of the final multivariate model: area of first enrollment, type of housing, mother education level, type of admission, payment scale, age, standardized academic performance, percentage of passed credits. Taking as reference students enrolled in formal sciences and engineering, students enrolled in humanities and social sciences have 43% (OR = 0.57, 95% IC: 0.53-0.62) less ODH, controlling at the same time by type of housing, mother education level, type of admission, payment scale, age, standardized academic performance and percentage of passed credits (or in general "the rest of variables" to mean those which are not mentioned). Students enrolled in these 2 areas represents more than 91% of the population in study. About type of housing, we observed that in the case of students who rent the ODH increases in 15% (OR = 1.15, 95% IC: 1.06-1.26) controlling at the same time by the rest of variables. As was analyzed before, the mother educational level is inversely proportional to the ODH, while the level is higher the ODH decreases. To have attended technical studies increases the ODH in 11% (OR = 1.11, 95% IC: 1.02-1.21), high school level increases the risk in 9% (OR = 1.09, 95% IC: 1.01-1.18), and primary school level increases the ODH in 27% (OR = 1.27, 95% IC: 1.04-1.56), all this scenarios in comparison with the college level and after controlling by the rest of variables. It was also found that when students obtained the admission by studying at the pre-university center their ODH decreases in 22% (OR = 0.78, 95% IC: 0.69-0.88) and if they were admitted by entrancing as member of the top third percent at the high school, then the ODH increases in 27% (OR = 1.27, 95% IC: 1.13-1.41), controlling at the same time by the rest of variables. The relation between payment scale and ODH keep being directly proportional, if the payment scale increases to the next category, then the ODH increases in 14% (OR = 1.14, 95% IC: 1.11-1.17), controlling by the rest of variables. The age is also directly proportional to the ODH, each year that a student is older, the ODH to leave increases in 3% (OR = 1.03, 95% IC: 1.01-1.05). The standardized academic performance index, as could be expected, has an inverse effect in the ODH, for each unit that it increases the ODH decreases in 8% (OR = 0.92, 95% IC: 0.92-0.93), controlling by the rest of variables. Finally, the percentage of passed credits performs an inverse relation with the ODH too, for each unit this variable increases the ODH decreases in 3% (OR = 0.97, 95% IC: 0.97-0.97), controlling by the rest of variables. By design we included the cohort effect in the estimation of the multivariate model. We found that while more recent is the cohort students are less prone to drop out (see Table (**??**)).

Table 3.5: Univariate regression analysis (1/3)

	Cox regression			Logistic regression		
Factor	HR	95% CI	p-value	OR	95% CI	p-value
Gender						
Female	1.00	–	–	1.00	–	–
Male	1.56	(1.48, 1.65)	<0.001	1.59	(1.50, 1.68)	<0.001
Marital status						
Single	1.00	–	–	1.00	–	–
Married and Others	1.81	(0.98, 3.37)	0.060	1.72	(0.84, 3.52)	0.136
Area of first enrollment						
Formal Sciences and Engineering	1.00	–	–	1.00	–	–
Architecture and Urbanism	0.99	(0.88, 1.11)	0.862	0.99	(0.87, 1.12)	0.835
Arts	0.81	(0.70, 0.93)	0.004	0.78	(0.67, 0.92)	0.003
Education	1.10	(0.89, 1.36)	0.356	1.15	(0.92, 1.44)	0.222
Humanities and Social Sciences	0.57	(0.54, 0.60)	<0.001	0.51	(0.48, 0.54)	<0.001
High school type						
Private / private religious	1.00	–	–	1.00	–	–
Public / public religious	1.07	(0.98, 1.16)	0.146	1.05	(0.96, 1.15)	0.315
Parochial	0.82	(0.72, 0.93)	0.002	0.77	(0.67, 0.89)	0.006
Others	0.79	(0.67, 0.93)	0.005	0.78	(0.65, 0.93)	<0.001
High school location						
Lima y Callao	1.00	–	–	1.00	–	–
Provinces	0.85	(0.79, 0.92)	<0.001	0.87	(0.80, 0.95)	0.001
Abroad	1.87	(1.12, 3.10)	0.016	2.02	(1.18, 3.49)	0.011
Type of housing						
Own	1.00	–	–	1.00	–	–
Renting	1.18	(1.10, 1.26)	<0.001	1.17	(1.08, 1.26)	<0.001
Living with relatives	1.05	(0.98, 1.12)	0.171	1.04	(0.97, 1.12)	0.301
Others	1.11	(0.98, 1.24)	0.100	1.09	(0.96, 1.24)	0.196
Indicator of mother alive						
Yes	1.00	–	–	1.00	–	–
No	1.34	(1.17, 1.53)	<0.001	1.36	(1.17, 1.58)	<0.001
Indicator of father alive						
Yes	1.00	–	–	1.00	–	–
No	1.13	(1.05, 1.21)	0.001	1.10	(1.02, 1.20)	0.017

*First showed value for each variable is the reference or pivot value

Table 3.6: Univariate regression analysis (2/3)

	Cox regression			Logistic regression		
Factor	HR	95% CI	p-value	OR	95% CI	p-value
Mother education level						
College education	1.00	–	–	1.00	–	–
Technical education	1.15	(1.07, 1.23)	<0.001	1.14	(1.06, 1.23)	<0.001
High school	1.37	(1.29, 1.46)	<0.001	1.34	(1.25, 1.43)	<0.001
Primary school	1.73	(1.48, 2.03)	<0.001	1.69	(1.42, 2.02)	<0.001
Father education level						
College education	1.00	–	–	1.00	–	–
Technical education	1.24	(1.15, 1.34)	<0.001	1.19	(1.09, 1.30)	<0.001
High school	1.39	(1.30, 1.49)	<0.001	1.34	(1.25, 1.45)	<0.001
Primary school	1.71	(1.39, 2.11)	<0.001	1.61	(1.27, 2.03)	<0.001
Number of relatives at the University						
Only student	1.00	–	–	1.00	–	–
Two	0.80	(0.70, 0.91)	<0.001	0.74	(0.64, 0.85)	<0.001
Three	0.74	(0.66, 0.82)	<0.001	0.68	(0.61, 0.77)	<0.001
Four	0.67	(0.61, 0.74)	<0.001	0.63	(0.57, 0.70)	<0.001
Five	0.99	(0.56, 1.76)	0.984	1.03	(0.56, 1.90)	0.933
Working						
No	1.00	–	–	1.00	–	–
Yes	1.67	(1.40, 2.00)	<0.001	1.45	(1.17, 1.80)	<0.001
Type of admission						
Entrance examination	1.00	–	–	1.00	–	–
Entrance examination - 1st attempt	0.68	(0.64, 0.72)	<0.001	0.71	(0.66, 0.76)	<0.001
Pre-university center	0.74	(0.67, 0.81)	<0.001	0.75	(0.67, 0.83)	<0.001
Entrance by top third percent	0.55	(0.51, 0.60)	<0.001	0.60	(0.54, 0.65)	<0.001
Special programs of admission	0.86	(0.76, 0.97)	0.011	0.97	(0.86, 1.10)	0.665
Payment scale **						
Payment scale	–	–		1.05	(1.02, 1.07)	<0.001
Request reclassification **						
No	–	–		1.00	–	–
Yes	–	–		0.78	(0.67, 0.92)	0.002

*First showed values for each variable was the reference or pivot value

**Time-varying variables

**For “education level” variable, each value means complete or incomplete studies at the respective level

Table 3.7: Univariate regression analysis (3/3)

	Cox regression			Logistic regression		
Factor	HR	95% CI	p-value	OR	95% CI	p-value
Type of benefit **						
None	–	–	–	1.00	–	–
Scholarship	–	–	–	0.56	(0.40, 0.78)	<0.001
Discount	–	–	–	0.72	(0.56, 0.91)	0.007
Loans	–	–	–	0.23	(0.14, 0.39)	<0.001
Others	–	–	–	10.23	(4.94, 21.19)	<0.001
Visits to health service **						
Number of visits	–	–	–	0.90	(0.88, 0.93)	<0.001
Age **						
Age	–	–	–	1.15	(1.13, 1.17)	<0.001
Academic performance index **						
CRAEST	–	–	–	0.85	(0.84, 0.85)	<0.001
Percentage of passed credits **						
Passed credits	–	–	–	0.96	(0.96, 0.96)	<0.001

*First showed values for each variable was the reference or pivot value

**Time-varying variables

Table 3.8: Multivariate logistic regression model (1/2)

	Logistic regression		
Factor	OR	95% CI	p-value
Area of first enrollment			
Formal Sciences and Engineering	1.00	–	–
Architecture and Urbanism	0.80	(0.70, 0.92)	<0.001
Arts	1.11	(0.92, 1,33)	0.270
Education	1.46	(1.12, 1.91)	<0.001
Humanities and Social Sciences	0.57	(0.53, 0.62)	<0.001
Type of housing			
Own	1.00	–	–
Renting	1.15	(1.06, 1.26)	<0.001
Living with relatives	1.13	(1.04, 1.23)	<0.001
Others	1.11	(0.96, 1.29)	0.150
Mother education level			
College education	1.00	–	–
Technical education	1.11	(1.02, 1.21)	0.020
High school	1.09	(1.01, 1.18)	0.030
Primary school	1.27	(1.04, 1.56)	0.020
Type of admission			
Entrance examination	1.00	–	–
Entrance examination - 1st attempt	1.11	(1.02, 1.21)	0.010
Pre-university center	0.78	(0.69, 0.88)	<0.001
Entrance by top third percent	1.27	(1.13, 1.41)	<0.001
Special programs of admission	1.96	(1.65, 2.34)	<0.001
Payment scale **			
Payment scale	1.14	(1.11, 1.17)	<0.001
Age **			
Age	1.03	(1.01, 1.05)	0.010
Standardized academic performance index **			
academic performance index	0.92	(0.92, 0.93)	<0.001
Percentage of passed credits **			
Passed credits	0.97	(0.97, 0.97)	<0.001

*First showed values for each variable was the reference or pivot value

**Time-varying variables

Table 3.9: Multivariate logistic regression model (2/2)

	Logistic regression		
Factor	OR	95% CI	p-value
Cohort***			
2004	1.00	–	–
2005	0.87	(0.77, 0.98)	0.020
2006	0.81	(0.72, 0.92)	<0.001
2007	0.89	(0.79, 1.00)	<0.001
2008	0.73	(0.64, 0.82)	<0.001
2009	0.69	(0.61, 0.78)	<0.001
2010	0.55	(0.48, 0.62)	<0.001
2011	0.28	(0.23, 0.34)	<0.001
2012	0.00	(0.00, ∞)	0.830

***In the analysis the cohort effect is included by design.

Chapter 4

Discussion and Future Research

"Providence has its appointed hour for everything.
We cannot command results, we can only strive".
Gandhi, 1939

4.1 Conclusions

In this paper, we reviewed the regression models for discrete data and applied them to the analysis of factors associated with dropout at a major private University in Perú. After fitting univariate and multivariate regression models, we found that the academic factors associated with student dropout were the area of study, modality of admission, standardized academic performance index, and percentage of passed credits. Important factors that could be associated with student's economy were the type of residence and payment scale. A personal features was also important: the mother education level, and in general the two features related with parents, indicators of whether or not are alive and education levels revealed to be more important if the mother is absent. The risk of dropout is directly proportional to the age of the student, it increases each year a student is older. We consider that the model's assumption of proportionality is so difficult to reach, at least taking as reference the comparison of hazard curves (or the logit of hazards curves), as we could see in the section of Kaplan-Meier estimators. However, in most of the observation period this assumption allows us to make the conclusion presented in section Results. One additional feedback we take, is that in this study, censored observations were taken as right-censoring cases, we can also assume this feature as interval-censored data given that we take measures at the end of one semesters and the real interruption of the studies can place in before or after each ending of semester. It implies to handle another likelihood function in order to calculate the respective estimators. It would be an interesting work to compare these two forms of estimations applied to a same case of study.

4.2 Future Research at Survival Analysis

- **Additive models**. The Nelson-Aalen estimator can be used to generate a cumulative hazard rate function (see equation (2.20)) and using this result the estimated survival function is $\hat{S}_j = e^{-\hat{\Lambda}_j}$.

- **Prediction**. We have focus on the problem of estimation. There is a remaining question on how to do model prediction under this setting. One approach is to study the time-dependent accuracy summaries based on time-specific versions of sensitivity and specificity calculated over risk sets (Heagerty, 2005).

- **Long-term survival**. In addition, the discrete-time survival model we have considered, we assumed that all individuals will eventually drop out the University. However, in this case of study, this is clearly unlikely. Current works attempt to adapt this discrete time survival model to consider a cure fraction (group of individuals that does not fail in the observation period)(Zhao and Zhou, 2008).

- **Bayesian survival methods**. Another approach to the estimate of survival functions is by using Bayesian methods. The current dropouts research problem can be performed by using this kind of techniques (Susarla and Van Ryzin, 1976).

- **Competing risk models**. There can be many causes of failure within a context. These models attempt to model the time of event occurrence under many conditions that can trigger the failure. The condition which occurs first will determine the survival time of the individual of interest. An example of this type of study and its merging with Bayesian estimation methods can be found on (Vallejos and Steel, 2014).

Appendix A

Theoretical notes and calculations

A.0.1 Probability density function of T expressed in terms of the hazard function

$$f_j = P(T=j) = \frac{\lambda_j}{(1-\lambda_j)} \prod_{k=1}^{j} (1-\lambda_k) = \frac{\lambda_j}{(1-\lambda_j)} S_j \tag{A.1}$$

To get last equation in (A.1) it considers 2.3

$$\lambda_j = \frac{P(T=j)}{P(T=j)+P(T>j)} = \frac{1}{1+\frac{S_j}{P(T=j)}} \tag{A.2}$$

and isolating the discrete density function

$$f_j = \frac{\lambda_j S_j}{1-\lambda_j} = \frac{\lambda_j}{(1-\lambda_j)} \prod_{k=1}^{j} (1-\lambda_k) = \lambda_j \prod_{k=1}^{j-1} (1-\lambda_k) \tag{A.3}$$

A.0.2 Hazard ratio interpretation

As examples, suppose there are two individuals a and a^* with vector of covariates $\boldsymbol{z}$ and $\boldsymbol{z}^*$ that only differs at the kth variable being $\boldsymbol{z}_{ka} = b$ for individual a and $\boldsymbol{z}_{ka^*} = b^*$ for individual a^*, then the hazard ratio will be $e^{(b-b^*)^t\beta}$ because

$$\begin{aligned} HR(t,\boldsymbol{z},\boldsymbol{z}^*) &= \frac{\lambda(t, z_1, ..., z_{k-1}, b, z_{k+1}, ..., z_p)}{\lambda(t, z_1, ..., z_{k-1}, b^*, z_{k+1}, ..., z_p)} \\ &= \frac{\lambda_0(t) e^{(\beta_1 z_1, ..., \beta_{k-1} z_{k-1}, b, \beta_{k+1} z_{k+1}, ..., \beta_p z_p)}}{\lambda_0(t) e^{(\beta_1 z_1, ..., \beta_{k-1} z_{k-1}, b^*, \beta_{k+1} z_{k+1}, ..., \beta_p z_p)}} \\ &= e^{(b-b^*)\beta_k} \end{aligned} \tag{A.4}$$

Notice that in the particular case that values in which differs individuals a and a^* were $z_k = 1$ for individual a and $z_k = 0$ for individual a^*, we have

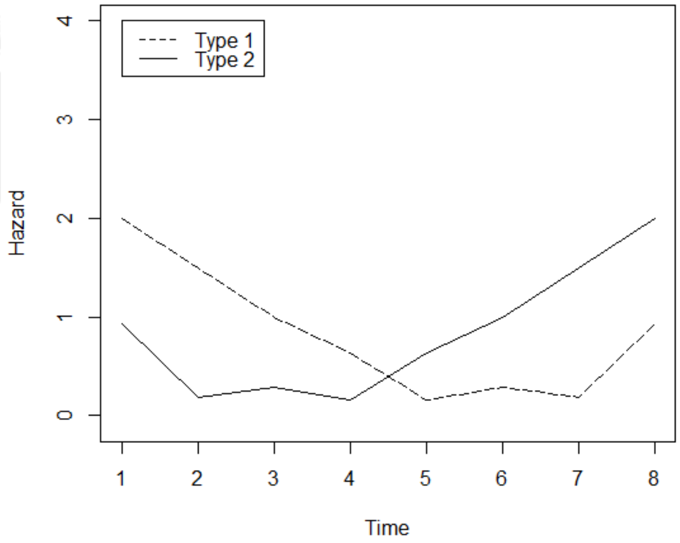
$$\begin{aligned} HR(t, \boldsymbol{z}, \boldsymbol{z}^*) &= \frac{\lambda(t, z_1, ..., z_{k-1}, 1, z_{k+1}, ..., z_p)}{\lambda(t, z_1, ..., z_{k-1}, 0, z_{k+1}, ..., z_p)} \\ &= \frac{\lambda_0(t)e^{(\beta_1 z_1, ..., \beta_{k-1} z_{k-1}, 1, \beta_{k+1} z_{k+1}, ..., \beta_p z_p)}}{\lambda_0(t)e^{(\beta_1 z_1, ..., \beta_{k-1} z_{k-1}, 0, \beta_{k+1} z_{k+1}, ..., \beta_p z_p)}} \\ &= e^{(1-0)\beta_k} \\ &= e^{\beta_k} \qquad (A.5) \end{aligned}$$

A.0.3 Proportional hazard assumption

The following example considers three hazard ratio values that shows the lack of proportionality along time. Let t_1,t_2 and t_3 be failure times and z a binary covariate, then if it needs to analyze the hazard ratio for individuals with $z = 1$ and $z^* = 0$ the function would be

$$HR(t, \boldsymbol{z}, \boldsymbol{z}^*) = \frac{\lambda(t, z = 1)}{\lambda(t, z = 0)} \qquad (A.6)$$

and if $HR(t_1, z, z^*) < 1$, $HR(t_2, z, z^*) = 1$, and $HR(t_3, z, z^*) > 1$, we might observe the following behavior:



then, the hazard ratio is not constant over time, the hazard functions are not parallel and even an intersection exists. Under this scenario this type of regression could not be the most suitable. Suggested options to deal with this scenario are: to split the analysis in two parts (before and after intersection time) and to estimate functions separately; or to include a time-dependent covariate in order to measure the interaction between the variable z and the time.

A.0.4 Likelihood ratio test

Table A.1: Likelihood ratio test

Model	DF	Chi-square	p-value
Regression models comparison			
Basic multivariate model (BMM) vs BMM + high school type	4	8.513219	0.074
Basic multivariate model (BMM) vs BMM + indicator or mother alive	1	0.1182135	0.731
Basic multivariate model (BMM) vs BMM + number of relatives	4	6.357758	0.174

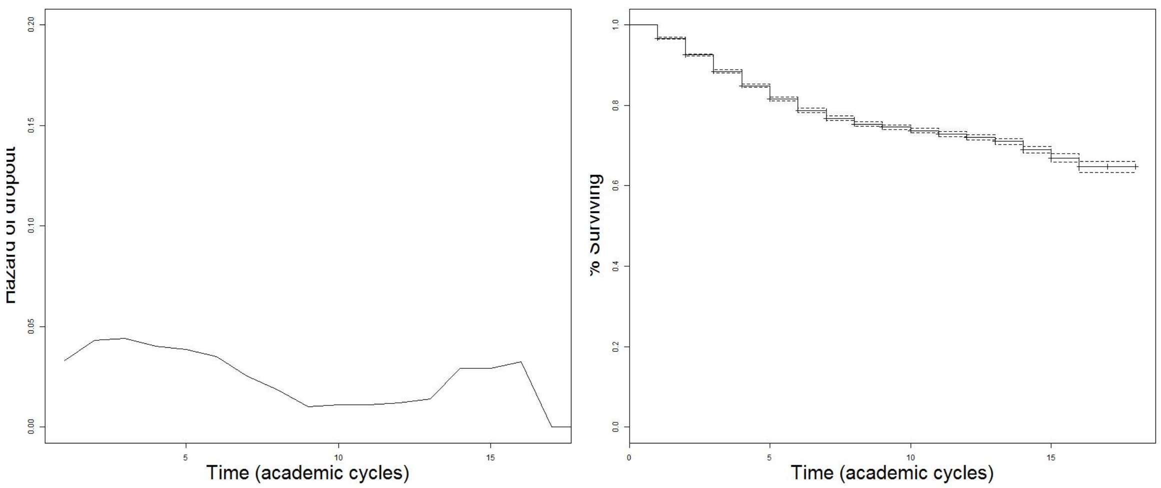
Appendix B

Graphical results

B.0.5 Kaplan-Meier estimators for the whole population

Figure B.1 shows the Kaplan-Meier estimator as the estimated hazard and survival functions for the entire population and its 95% (Wald) confidence intervals. The graphic shows also ticks in intervals where there are censored observations.

Figure B.1: Kaplan-Meier estimator for the entire population



B.0.6 Estimated hazard functions or logit(hazard) functions for relevant covariates

Figure B.2: Hazard for entire population by first enrolled area

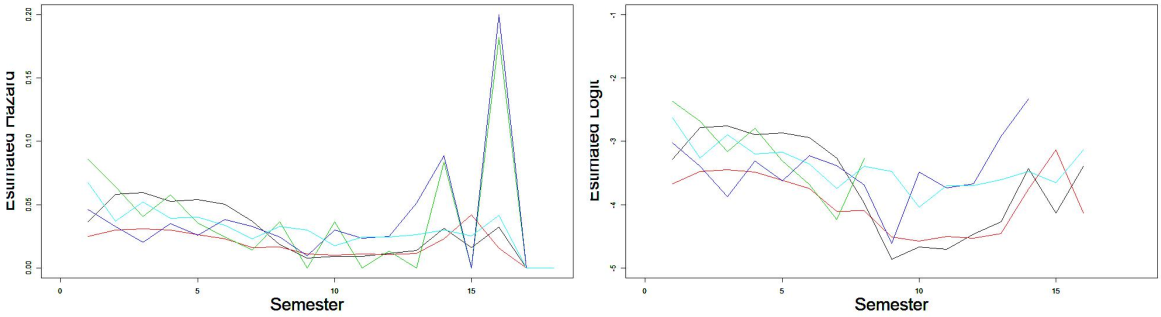


Figure B.3: Hazard for entire population by residence type

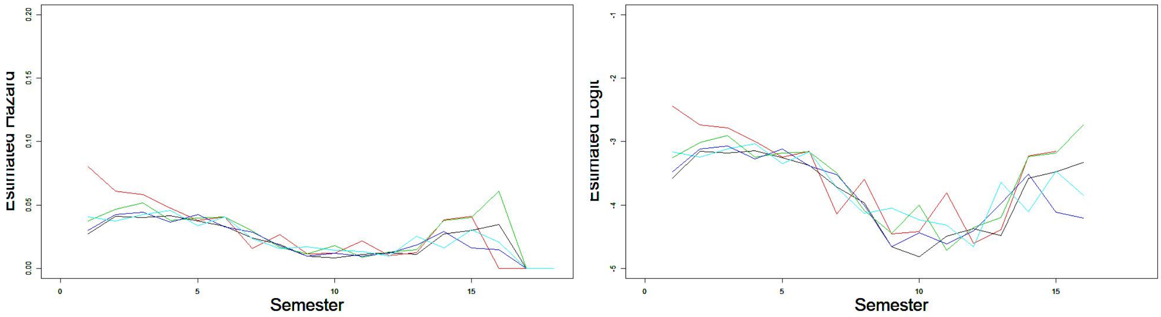


Figure B.4: Hazard for entire population by mother level education

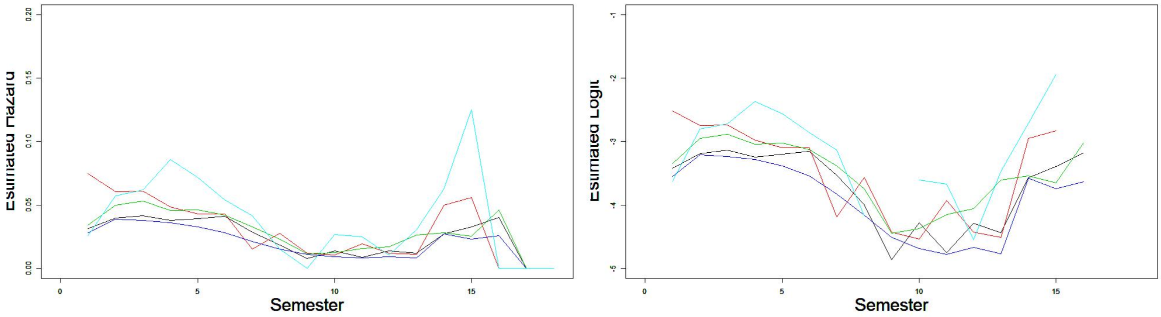
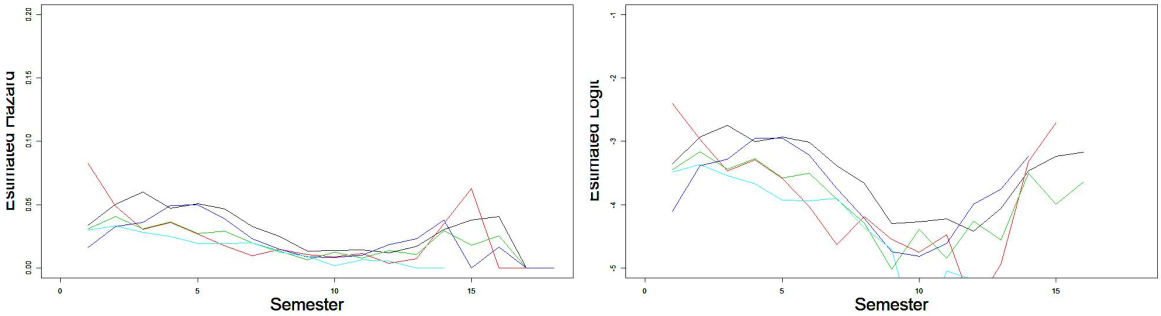


Figure B.5: Hazard for entire population by admission type



Appendix C

Source code

C.0.7 Simulation of proportionality assumption

```
#COX MODEL
par(mfrow=c(1,1))

#Simulating HAZARD, in black
logit1 = c(runif(5),1,1.5,2) #random values for simulation
logit2 = logit1 + 1 #parallel function
plot(logit1,type="l",ylim=c(-0.1,4),main="Hazard",xlab="Time",ylab="Hazard")
lines(logit2)

#LOGISTIC MODEL
par(mfrow=c(1,3))

#Simulating LOGIT(HAZARD), in black
logit1 = c(runif(5),1,1.5,2) #random values for simulation
logit2 = logit1 + 1 #parallel function
plot(logit1,type="l",ylim=c(-0.1,4),main="Logit",xlab="Time",ylab="Logit")
lines(logit2)

#Calculating ODDS, in blue
odds1 = exp(logit1)
odds2 = exp(logit2)
plot(odds1,type="l",col="blue",ylim=c(-0.1,20),main="Odds",xlab="Time",ylab="Odds")
lines(odds2,col="blue")

#Calculating HAZARD, in red
hazard1 = 1/(1+exp(-logit1))
hazard2 = 1/(1+exp(-logit2))
plot(hazard1,type="l",col="red",ylim=c(-0.1,1.5),main="Hazard",xlab="Time",ylab="Hazard")
lines(hazard2,col="red")
```

C.0.8 Processing source code in R 3.0

```
rm(list=ls(all=TRUE))
library(splines); library(survival); library(foreign); library(xtable); library(plyr); library(stringr)

#FUNCTION
trim <- function (x) gsub("^\\s+|\\s+$", "", x)
par(mar=c(4,4,0.5,0.5)); #par(mfrow=c(1,3))
```

```
#01 - DATA LOADING
dir = "D:/Miguel Pebes/02 - PUCP - Maestría en Estadística/13 Ciclo4 - Seminario de Tesis/
MPEBES_Sustentacion/R App/"
graph_dir = "D:/Miguel Pebes/02 - PUCP - Maestría en Estadística/13 Ciclo4 - Seminario de Tesis/
MPEBES_Sustentacion/graphics/"
datos <- read.csv(paste(dir,"archivo 1-alumno-final.csv",sep="")); names(datos)
datos2<- read.csv(paste(dir,"archivo 2a-alumno-cicmat.csv",sep="")); names(datos2)
datos3<- read.csv(paste(dir,"archivo 2b-alumno-cicegre-sinrep.csv",sep=""))
datos2_preprocesado <- read.csv(paste(dir,"archivo 2a-alumno-cicmat-preprocesado.csv",sep=""))

#02 - FILTERING DATABASE
#Selecting information from 2004 on out and not including summer schedules

datos = subset(datos,datos$d_cic_1rma!='2002-1' & datos$d_cic_1rma!='2002-2' & datos$d_cic_1rma!='2003-1'
& datos$d_cic_1rma!='2003-2'); table(datos$d_cic_1rma,exclude=NULL)
datos2 = subset(datos2,datos2$cicmat!='2002-1' & datos2$cicmat!='2002-2' & datos2$cicmat!='2003-0'
& datos2$cicmat!='2003-1' & datos2$cicmat!='2003-2' & datos2$cicmat!='2004-0')
table(datos2$cicmat,exclude=NULL)
datos2_preprocesado = subset(datos2_preprocesado,datos2_preprocesado$cicmat!='2002-1'
& datos2_preprocesado$cicmat!='2002-2' &
datos2_preprocesado$cicmat!='2003-1' & datos2_preprocesado$cicmat!='2003-2'
& substr(datos2_preprocesado$cicmat,5,6)!='-0')

#02 - PREPROCESSING
#Grouping/creating categories
#
table(datos$sexo)
datos$sexo=relevel(datos$sexo,ref="FEMENINO")
#
#table(datos$d_estado_civil_ingreso)
datos$d_estado_civil_ingreso = as.character(datos$d_estado_civil_ingreso)
vector = c("CASADO(A)","CONVIVIENTE","RELIGIOSO(A)","SEPARADO(A)")
datos$d_estado_civil_ingreso[which(datos$d_estado_civil_ingreso %in% vector)] = "OTRO"
table(datos$d_estado_civil_ingreso,exclude=NULL)
datos$d_estado_civil_ingreso=factor(datos$d_estado_civil_ingreso)
datos$d_estado_civil_ingreso=relevel(datos$d_estado_civil_ingreso,ref="SOLTERO(A)")
#
table(datos$d_area_1rma)
datos$d_area_1rma=relevel(datos$d_area_1rma,ref="CIENCIAS")
#
#table(datos$d_tipocol)
datos$d_tipocol = as.character(datos$d_tipocol)
datos$d_tipocol[datos$d_tipocol == "NACIONAL RELIGIOSO"] = "NACIONAL"
vector = c("PARTICULAR LAICO","PARTICULAR RELIGIOSO")
datos$d_tipocol[which(datos$d_tipocol %in% vector)] = "PARTICULAR"
vector = c("FE Y ALEGRIA","FUERZAS ARMADAS")
datos$d_tipocol[which(datos$d_tipocol %in% vector)] = "OTRO"
table(datos$d_tipocol,exclude=NULL)
datos$d_tipocol=factor(datos$d_tipocol)
datos$d_tipocol=relevel(datos$d_tipocol,ref="PARTICULAR")
#
```

```
#table(datos$d_dptocol)
datos$d_dptocol = as.character(datos$d_dptocol)
vector = c('AMAZONAS','ANCASH','APURIMAC','AREQUIPA','AYACUCHO','CAJAMARCA','CUSCO','HUANCAVELICA',
'HUANUCO','ICA','JUNIN','LA LIBERTAD','LAMBAYEQUE','LORETO','MADRE DE DIOS','MOQUEGUA','PASCO','PIURA',
'PUNO','SAN MARTIN','TACNA','TUMBES','UCAYALI')
datos$d_dptocol[which(datos$d_dptocol %in% vector)] = "PROVINCIA"
vector = c('LIMA','CALLAO')
datos$d_dptocol[which(datos$d_dptocol %in% vector)] = "LIMA Y CALLAO"
table(datos$d_dptocol,exclude=NULL)
datos$d_dptocol=factor(datos$d_dptocol)
datos$d_dptocol=relevel(datos$d_dptocol,ref="LIMA Y CALLAO")
#
#table(datos$d_tipo_residencia)
datos$d_tipo_residencia  = as.character(datos$d_tipo_residencia)
vector = c('ALQUILER-VENTA')
datos$d_tipo_residencia[which(datos$d_tipo_residencia %in% vector)] = "ALQUILADA"
vector = c('EN CALIDAD DE USO','GUARDIANÍA','INVASIÓN U OCUPACIÓN PRECARIA')
datos$d_tipo_residencia[which(datos$d_tipo_residencia %in% vector)] = "OTRO"
table(datos$d_tipo_residencia,exclude=NULL)
datos$d_tipo_residencia=factor(datos$d_tipo_residencia)
datos$d_tipo_residencia=relevel(datos$d_tipo_residencia,ref="PROPIA")
#
table(datos$i_mad_viva,exclude=NULL)
datos$i_mad_viva = ifelse(is.na(datos$i_mad_viva),"1",datos$i_mad_viva)
datos$i_mad_viva=factor(datos$i_mad_viva)
datos$i_mad_viva=relevel(datos$i_mad_viva,ref="1")
#
table(datos$i_pad_vivo,exclude=NULL)
datos$i_pad_vivo = ifelse(is.na(datos$i_pad_vivo),"1",datos$i_pad_vivo)
datos$i_pad_vivo=factor(datos$i_pad_vivo)
datos$i_pad_vivo=relevel(datos$i_pad_vivo,ref="1")
#
#table(datos$d_gradoinstru_madre)
datos$d_gradoinstru_madre  = as.character(datos$d_gradoinstru_madre)
vector = c('POST-GRADO','UNIVERSIDAD COMPLETA','UNIVERSIDAD INCOMPLETA')
datos$d_gradoinstru_madre[which(datos$d_gradoinstru_madre %in% vector)] = "UNIVERSITARIA"
vector = c('TECNICA COMPLETA','TECNICA INCOMPLETA')
datos$d_gradoinstru_madre[which(datos$d_gradoinstru_madre %in% vector)] = "TECNICA"
vector = c('SECUNDARIA COMPLETA','SECUNDARIA INCOMPLETA')
datos$d_gradoinstru_madre[which(datos$d_gradoinstru_madre %in% vector)] = "SECUNDARIA"
vector = c('PRIMARIA COMPLETA','PRIMARIA INCOMPLETA')
datos$d_gradoinstru_madre[which(datos$d_gradoinstru_madre %in% vector)] = "PRIMARIA"
table(datos$d_gradoinstru_madre,exclude=NULL)
datos$d_gradoinstru_madre=factor(datos$d_gradoinstru_madre)
datos$d_gradoinstru_madre=relevel(datos$d_gradoinstru_madre,ref="UNIVERSITARIA")
#
#table(datos$d_gradoinstru_padre)
datos$d_gradoinstru_padre  = as.character(datos$d_gradoinstru_padre)
vector = c('POST-GRADO','UNIVERSIDAD COMPLETA','UNIVERSIDAD INCOMPLETA')
datos$d_gradoinstru_padre[which(datos$d_gradoinstru_padre %in% vector)] = "UNIVERSITARIA"
vector = c('TECNICA COMPLETA','TECNICA INCOMPLETA')
datos$d_gradoinstru_padre[which(datos$d_gradoinstru_padre %in% vector)] = "TECNICA"
```

```
vector = c(’SECUNDARIA COMPLETA’,’SECUNDARIA INCOMPLETA’)
datos$d_gradoinstru_padre[which(datos$d_gradoinstru_padre %in% vector)] = "SECUNDARIA"
vector = c(’PRIMARIA COMPLETA’,’PRIMARIA INCOMPLETA’)
datos$d_gradoinstru_padre[which(datos$d_gradoinstru_padre %in% vector)] = "PRIMARIA"
table(datos$d_gradoinstru_padre,exclude=NULL)
datos$d_gradoinstru_padre=factor(datos$d_gradoinstru_padre)
datos$d_gradoinstru_padre=relevel(datos$d_gradoinstru_padre,ref="UNIVERSITARIA")
#
table(datos$nintegrantes_familia,exclude=NULL)
datos$nintegrantes_familia = ifelse(is.na(datos$nintegrantes_familia),"1",datos$nintegrantes_familia)
datos$nintegrantes_familia=factor(datos$nintegrantes_familia)
datos$nintegrantes_familia=relevel(datos$nintegrantes_familia,ref="1")
#
table(datos$i_trabaja_cp,exclude=NULL)
datos$i_trabaja_cp=factor(datos$i_trabaja_cp)
datos$i_trabaja_cp=relevel(datos$i_trabaja_cp,ref="0")
#
#table(datos$D_MODALIDAD)
datos$D_MODALIDAD = as.character(datos$D_MODALIDAD)
vector = c(’DIPLOMAS DE BACHILLERATO’,’EXAMEN DE INGRESO ORDINARIO Y POR EXONERACION’,
’INGRESO BICENTENARIO DE LA INDEPENDENCIA DEL PERU’,
’INGRESO POR CONVENIO COLEGIOS FE Y ALEGRIA’,’INGRESO R.P. JORGE DINTILHAC SS.CC.’,
’PROGRAMA DE EXCELENCIA ESCOLAR’)
datos$D_MODALIDAD[which(datos$D_MODALIDAD %in% vector)] = "ESPECIAL"
table(datos$D_MODALIDAD,exclude=NULL)
datos$D_MODALIDAD=factor(datos$D_MODALIDAD)
datos$D_MODALIDAD=relevel(datos$D_MODALIDAD,ref="EVALUACION DEL TALENTO Y EXONERADOS")
#
table(datos$d_cic_1rma,exclude=NULL)
datos$d_cic_1rma=relevel(datos$d_cic_1rma,ref="2004-1")
table(substr(datos$d_cic_1rma,5,6),exclude=NULL) #Validation only semesters I and II, not summer cycles

#Inputs Generation

temp1 = datos2[substr(datos2$cicmat,5,6)!="-0",]
temp2 = temp1[,c("numsec","cicmat")]
temp3 = cbind(temp2[-2], model.matrix(~0+cicmat,temp2))
#temp4 = head(temp3,2500)
temp4 = temp3
temp5 = do.call("rbind",as.list(by(temp4,list(numsec=temp4$numsec),
function(x){y<-subset(x,select=-numsec);apply(y,2,sum)})))
#temp6 = apply(temp5,1,paste,collapse=" ") #Here we have a life pattern for each individual

temp5 = data.frame(temp5)
temp5.5 = temp5[,names(temp5)[which(substr(names(temp5),11,12) %in% c(".1",".2"))]]
#temp6 = data.frame(str_replace_all(string=apply(temp5.5,1,paste,collapse=" "), pattern=" ", repl=""))
temp6 = str_replace_all(string=apply(temp5.5,1,paste,collapse=" "), pattern=" ", repl="")
temp7 = data.frame(unique(temp4$numsec)); colnames(temp7) = c("numsec"); temp7
temp8 = datos3[,c("numsec","cicegr")] #identifying individuals that finish
temp9 = merge(temp7,temp8,by="numsec",all.x=TRUE)
#temp10= cbind(temp9,str_replace_all(string=temp6, pattern=" ", repl=""),
data.frame(regexpr("100",str_replace_all(string=temp6, pattern=" ", repl=""))[1:nrow(temp9)]))
```

```
temp10= cbind(temp9,str_replace_all(string=temp6, pattern=" ", repl=""),
data.frame(regexpr("100",str_replace_all(string=temp6, pattern=" ", repl=""))[1:nrow(temp9)]),
data.frame(regexpr("100",str_replace_all(string=temp6, pattern=" ", repl=""))[1:nrow(temp9)])-
data.frame(regexpr("1",str_replace_all(string=temp6, pattern=" ", repl=""))[1:nrow(temp9)])+1
)
colnames(temp10) = c("numsec","cicegr","pattern","pospattern","time") #; temp10
temp10$egr = "0"
temp10$egr[which(temp10$numsec %in% temp10$numsec[!is.na(temp10$cicegr)])] = "1" #; temp10
#temp10$time = ifelse(temp10$time!=-1,temp10$time,ifelse(substr(temp10$pattern,21,22)==’10’,
21,22)) #; temp10
temp10$time = ifelse(temp10$time>=0,temp10$time,ifelse(substr(temp10$pattern,21,22)==’10’,21,22)
+temp10$time+1) #; temp10

temp10$censor = "1"
temp10$censor = ifelse(temp10$egr==1,0,temp10$censor)#; temp10
#temp10$censor = ifelse(temp10$pospattern==-1,0,temp10$censor); temp10
temp10$censor = ifelse(temp10$pospattern<0,0,temp10$censor)#; temp10

temp11 = temp10[,c("numsec","pattern","time","censor","egr")]#; temp11 #SAL Y ROSAS
#temp11 = temp10[,c("numsec","pattern","time","censor")]#; temp11

temp12 = datos[,c(’numsec’,’sexo’,’d_estado_civil_ingreso’,’d_area_1rma’,’d_tipocol’,’d_dptocol’,
’d_tipo_residencia’,’i_mad_viva’,’i_pad_vivo’,’d_gradoinstru_madre’,
’d_gradoinstru_padre’,’nintegrantes_familia’,’i_trabaja_cp’,’D_MODALIDAD’,’d_cic_1rma’)]

temp13 = merge(temp11,temp12,by="numsec")
temp13 = merge(temp11,temp12,by="numsec",all.y=TRUE); head(temp13,10); length(temp13); nrow(temp13)
temp13$time = as.numeric(temp13$time)
temp13$censor = as.numeric(temp13$censor)

#KAPLAN-MEIER ESTIMATIONS FOR POPULATION
#Graphic configuration
mp = paste(graph_dir,"01_KaplanMeierEstimation.png",sep = "")
mypath <- file.path(mp)
#jpeg(file=mypath, width = 2300, height = 1400)
jpeg(file=mypath, width = 1800, height = 800)
par(mfrow=c(1,2))

datos = temp13

#Hazard function
mod1 <- survfit(Surv(time,censor)~1,data=datos)
h0   <- mod1$n.event/mod1$n.risk
plot(mod1$time,h0,type="l",ylab="Hazard of dropout",xlab="Time (academic cycles)",ylim=c(0,0.2),
xlim=c(1,17),cex.lab=2.5)

#Survival function
#plot(mod1$time,mod1$surv,type="l",ylab="% Surviving",xlab="Time (academic cycles)",ylim=c(0,1),
xlim=c(1,17),cex.lab=2.5)
plot(survfit(Surv(time,censor)~1,data=datos),conf.int=TRUE,xlab="Time (academic cycles)",
ylab="% Surviving",cex.lab=2.5)
```

```
tab1 <- cbind(time=mod1$time,nriesgo=mod1$n.risk,evento=mod1$n.event,Riesgo=h0,Supevivencia=mod1$surv)
tab1 <- cbind(time=mod1$time,nriesgo=mod1$n.risk,evento=mod1$n.event,Riesgo=h0,Supevivencia=mod1$surv,
CI95low=mod1$lower,CI95up=mod1$upper)
print(xtable(tab1,digits=c(0,0,0,0,3,3,3,3)),include.rownames=FALSE)
summary(mod1)

dev.off()

#INVARIANT DATA:
#gender,area,typeschool,schoolloc,typehouse,status,motheraliv,fatheraliv,
#edlevmoth,edlevfath,numrelativ,working,tadmission,cohort
names(datos)

vector_invariant_data = c("sexo","d_area_1rma","d_tipocol","d_dptocol","d_tipo_residencia",
"d_estado_civil_ingreso","i_mad_viva","i_pad_vivo","d_gradoinstru_madre","d_gradoinstru_padre",
"nintegrantes_familia","i_trabaja_cp","D_MODALIDAD","d_cic_1rma")

#PROPORTIONALITY TEST
#---

#LOGRANK TEST
#defining vector of "survdiff" objects
univariate_logrank  = vector("list",length(vector_invariant_data))
for (i in 1:length(vector_invariant_data)){
#logrank
formula = paste("survdiff( Surv(time, censor)","~",vector_invariant_data[i]," , rho = 0,
data=datos)",sep="")
##rho = 0 prueba de Log Rank (por defecto), rho = 1 es la prueba de Gehan-Wilcoxon
logrankfit = eval(parse(text=formula)); print(formula)
univariate_logrank[[i]] = logrankfit
#degrees of freedom
formula = paste("unique(datos$",vector_invariant_data[i],")",sep="")
unique = eval(parse(text=formula)); print(formula)
num_values = length(unique); print(num_values-1) #Showing number of different values minus 1
#p-value
p = 1-pchisq(logrankfit$chisq,df=num_values-1); print(p)
}

#univariate_logrank[[1]]; univariate_logrank[[1]]$chisq;
p = 1-pchisq(univariate_logrank[[1]]$chisq,df=1); p
#---

#COX REGRESSION
#defining vectors of coxph objects
univariate_cox_regression = vector("list",length(vector_invariant_data))
for (i in 1:length(vector_invariant_data)){
#Bivariate regression
formula = paste("coxph(Surv(time,censor)~",vector_invariant_data[i],", data=datos)",sep="")
coxfit = eval(parse(text=formula)); print(formula)
univariate_cox_regression[[i]] = coxfit
summary(univariate_cox_regression[[i]])
}
```

```
#summary(univariate_cox_regression[[1]]);
#exp(summary(univariate_cox_regression[[1]])$coefficients[,1]);
#exp(confint(univariate_cox_regression[[1]]));
#---

#for (i in 1:length(vector_invariant_data)){
# par(mfrow=c(1,2)) #to draw
# formula = paste("unique(datos$",vector_invariant_data[i],")",sep="")
# print(formula)
# unique = eval(parse(text=formula)); #print(formula)
# unique = na.omit(unique)
# num_values = length(unique)
# col_vector = c(1:num_values) #to draw
#
# for (j in 1:num_values){
# print(paste(">",unique[j]))
# }
#}

#HAZARD AND SURVIVAL FUNCTIONS
for (i in 1:length(vector_invariant_data)){
formula = paste("unique(datos$",vector_invariant_data[i],")",sep="")
print(formula)
unique = eval(parse(text=formula)); #print(formula)
unique = na.omit(unique)
num_values = length(unique)
col_vector = c(1:num_values) #to draw

#defining vectors of "survfit" objects
hazardfit = vector("list",length(unique))

for (j in 1:num_values){
#print(paste(">",unique[j]))
#Estimation by groups
#formula_2ndlev = paste("survfit( Surv(time, censor)˜ 1, conf.type=’none’,
subset=(",vector_invariant_data[i],"==’",unique[j],"’ & ",vector_invariant_data[i],"!=’NA’),data=datos)",
sep="")
formula_2ndlev = paste("survfit( Surv(time, censor)˜ 1, conf.type=’none’,
subset=(",vector_invariant_data[i],"==’",unique[j],"’),data=datos)",sep="")
hazardfit[[j]] = eval(parse(text=formula_2ndlev)); print(formula_2ndlev)
}

#Graphic configuration
mp = paste(graph_dir,"01_KaplanMeierEstimation_byVariables_",vector_invariant_data[i],".png",sep = "")
mypath <- file.path(mp)
jpeg(file=mypath, width = 2000, height = 400) #2100,700
par(mfrow=c(1,3)) #to draw #c(1,2)

#drawing hazard
for (j in 1:num_values){
h <-hazardfit[[j]]$n.event/hazardfit[[j]]$n.risk
plot(hazardfit[[j]]$time,h,type="l",ylab="Estimated Hazard probability",xlab="Semester",ylim=c(0.0,0.2),
xlim=c(0,18),cex.lab=2.5,col=col_vector[j])
par(new=T)
}
par(new=F)

#drawing logit
for (j in 1:num_values){
h <-hazardfit[[j]]$n.event/hazardfit[[j]]$n.risk
plot(hazardfit[[j]]$time,log(h/(1-h)),type="l",ylab="Estimated Logit",xlab="Semester",ylim=c(-5,-1),
xlim=c(0,18),cex.lab=2.5,col=col_vector[j])
par(new=T)
}
par(new=F)

#drawing survival
for (j in 1:num_values){
plot(hazardfit[[j]]$time, hazardfit[[j]]$surv,type="l",ylab="Estimated Survival
Function",xlab="Semester",ylim=c(0,1),xlim=c(0,18),cex.lab=2.5,col=col_vector[j])
par(new=T)
}
par(new=F)
dev.off()
}
#coxfit <- coxph(Surv(time,censor)~group, data=datos); summary(coxfit)
#logrankfit = survdiff( Surv(time, censor)~ group, rho = 0, data=datos)
#p = 1-pchisq(logrankfit$chisq,df=1); p
#---

#TIME VARYING VARIABLES: use objects "datos" and "datos_timevarying"
datos_timevarying = datos2_preprocesado
datos = merge(datos_timevarying,datos,by="numsec")
temp14 = data.frame(datos); vcount = data.frame(table(temp14$numsec))
names(vcount) = c("numsec","frec")
datos = merge(datos,vcount,by="numsec")
datos$int <- as.numeric(unlist(apply(array(1:nrow(vcount)),1,function(x) c(1:vcount$frec[x]))))
#si es evento y ...
datos$event = ifelse(datos$censor==1,ifelse(datos$time==datos$int,1,0),0)
datos = datos[datos$int<=datos$time,]
#First 30 lines are shown. Check with the pattern of semesters
head(datos,15)

porc_cred_aprob = datos$crapdox/datos$craplex
porc_cred_aprob = porc_cred_aprob
datos = cbind(datos,porc_cred_aprob)
head(datos[,c("crapdox","craplex","porc_cred_aprob")],50)
nivel_aprob = datos$porc_cred_aprob * 100
datos = cbind(datos,nivel_aprob)
head(datos[,c("crapdox","craplex","porc_cred_aprob","nivel_aprob")],50)

#Example time varying for variable GENDER
mod_glm <- glm(event~factor(int)+sexo,family="binomial",data=datos)
summary(mod_glm)
```

```
sum.coef<-summary(mod_glm)$coef; est<-exp(sum.coef[,1])
upper.ci<-exp(sum.coef[,1]+qnorm(0.975)*sum.coef[,2])
lower.ci<-exp(sum.coef[,1]-qnorm(0.975)*sum.coef[,2])
xtable(cbind(est,lower.ci,upper.ci))

#Massive logistic regression
#Logistic model estimation for all covariates fixed and time-varying
univariate_logistic_regression = vector("list",length(vector_invariant_data))
for (i in 1:length(vector_invariant_data)){
#Logistic regression
formula = paste("mod_glm <- glm(event˜factor(int)+",vector_invariant_data[i],
",family=’binomial’,data=datos)",sep="")
logisticfit = eval(parse(text=formula)); print(formula)
univariate_logistic_regression[[i]] = logisticfit
summary(univariate_logistic_regression[[i]])
sum.coef<-summary(univariate_logistic_regression[[i]])$coef; est<-exp(sum.coef[,1])
upper.ci<-exp(sum.coef[,1]+qnorm(0.975)*sum.coef[,2])
lower.ci<-exp(sum.coef[,1]-qnorm(0.975)*sum.coef[,2])
}
#summary(univariate_logistic_regression[[1]]);
summary(univariate_logistic_regression[[1]]);
sum.coef<-summary(univariate_logistic_regression[[1]])$coef; est<-exp(sum.coef[,1])
upper.ci<-exp(sum.coef[,1]+qnorm(0.975)*sum.coef[,2])
lower.ci<-exp(sum.coef[,1]-qnorm(0.975)*sum.coef[,2])
xtable(cbind(est,lower.ci,upper.ci))

#MULTIPLE REGRESSION

mod_glm_mult <- glm(event˜factor(int)+sexo+d_area_1rma+d_tipocol+d_dptocol+d_tipo_residencia+
d_estado_civil_ingreso+d_estado_civil_ingreso+i_mad_viva+i_pad_vivo+d_gradoinstru_madre+
d_gradoinstru_padre+nintegrantes_familia+i_trabaja_cp+D_MODALIDAD+d_cic_1rma+escala+askrecateg+
tipobenef+ncitas+edad+craest,family=’binomial’,data=datos)

#Baseline: mod_glm_mult2$coefficients[1:18]
# vectorALFAs = mod_glm_mult$coefficients[1:18]
# vectorDsRowByRow = model.matrix(˜factor(int)-1,data=datos)
# vectorDs <- apply(vectorDsRowByRow,2,sum)
#mod_glm_mult$coefficients[1:18]
#summary.glm(mod_glm_mult)$coef[1:18,1] #same result former line
#um.coef<-summary(mod_glm_mult)$coef; est<-exp(sum.coef[,1])
#upper.ci<-exp(sum.coef[,1]+qnorm(0.975)*sum.coef[,2])
#lower.ci<-exp(sum.coef[,1]-qnorm(0.975)*sum.coef[,2])
#xtable(cbind(est,lower.ci,upper.ci))
#plot(exp(summary.glm(mod_glm_mult)$coef[1:18,1]))
#head(model.matrix(˜factor(int)-1,data=datos),20)
#row.sums <- apply(head(model.matrix(˜factor(int)-1,data=datos),20) ,2,sum)
#vectorDsRowByRow = model.matrix(˜factor(int)-1,data=datos)
#vectorDs <- apply(vectorDsRowByRow,2,sum)
#vectorALFAs = summary.glm(mod_glm_mult)$coef[1:18,1]
#linear_component <- apply( rbind(vectorDs,vectorALFAs) , 2, prod)

#Multivariate model - var 25% significance
```

```
#MULTIVARIANT MODEL
mod_glm_mult <- glm(event~factor(int)+d_area_1rma+d_tipo_residencia+d_gradoinstru_madre
+D_MODALIDAD+escala+edad+craest+nivel_aprob,family=’binomial’,data=datos)
sum.coef<-summary(mod_glm_mult)$coef; est<-exp(sum.coef[,1])
upper.ci<-exp(sum.coef[,1]+qnorm(0.975)*sum.coef[,2])
lower.ci<-exp(sum.coef[,1]-qnorm(0.975)*sum.coef[,2])
xtable(cbind(est,lower.ci,upper.ci,summary(mod_glm_mult)$coef[,4]))

#MULTIVARIANT EVALUATION

mod_glm_mult_1 <- glm(event~factor(int)+d_area_1rma+d_tipo_residencia+d_gradoinstru_madre
+D_MODALIDAD+escala+edad+craest+nivel_aprob+d_tipocol,family=’binomial’,data=datos)

datos$i_mad_viva = ifelse(is.na(datos$i_mad_viva),"1",datos$i_mad_viva)

mod_glm_mult_2 <- glm(event~factor(int)+d_area_1rma+d_tipo_residencia+d_gradoinstru_madre
+D_MODALIDAD+escala+edad+craest+nivel_aprob+i_mad_viva,family=’binomial’,data=datos)

mod_glm_mult_3 <- glm(event~factor(int)+d_area_1rma+d_tipo_residencia+d_gradoinstru_madre
+D_MODALIDAD+escala+edad+craest+nivel_aprob+nintegrantes_familia,family=’binomial’,data=datos)

#Likelihood ratio test for variables (only test)

library(epicalc)
lrtest(mod_glm_mult,mod_glm_mult_1)
lrtest(mod_glm_mult,mod_glm_mult_2)
lrtest(mod_glm_mult,mod_glm_mult_3)

ci = confint(mod_glm_mult)
exp(ci)

###Controling by ANNO###
datos$anno = substr(datos$d_cic_1rma,1,4)
mod_glm_mult_extra <- glm(event~factor(int)+d_area_1rma+d_tipo_residencia+d_gradoinstru_madre
+D_MODALIDAD+escala+edad+craest+nivel_aprob+anno,family=’binomial’,data=datos)
#Coeff summary(mod_glm_mult_extra)$coef[,1]
#OR exp(summary(mod_glm_mult_extra)$coef[,1])
#P-value summary(mod_glm_mult_extra)$coef[,4]

sum.coef<-summary(mod_glm_mult_extra)$coef; est<-exp(sum.coef[,1])
upper.ci<-exp(sum.coef[,1]+qnorm(0.975)*sum.coef[,2])
lower.ci<-exp(sum.coef[,1]-qnorm(0.975)*sum.coef[,2])
xtable(cbind(est,lower.ci,upper.ci,summary(mod_glm_mult_extra)$coef[,4]))
##########
```

Bibliography

Aalen, O. (1976). Nonparametric inference in connection with multiple decrement models, *The Scandinavian Journal of Statistics* **3**: 15–27.

Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle, *B. N. Petrov and F. Csaki, eds., 2nd International Symposium on Information Theory, Akademia Kiado, Budapest* pp. 267–281.

Allison, P. D. (1982). Discrete-time methods for the analysis of event histories, *Sociological Methodology* **13**: 61–98.

Aranda-Ordaz, F. J. (1983). An extension of the proportional-hazards model for grouped data, *Biometrics* **39**(1): 109–117.

Arulampalam, W., Naylor, R. A. and Smith, J. P. (2004). A hazard model of the probability of medical school, *Journal of the Royal Statistical Society* (167): 157–178.

Borges, R. E. (2005). Survival data analysis of patients with peritoneal dialysis, *Revista Colombiana de Estadistica* .

Bray, S., Gedeon, J., Hadi, A., Kotb, A., Rahman, T., Sarwar, E., Savelyeva, A., Sevigny, M., Bakanda, C., Birungi, J., Chan, K., Yaya, S., Deonandan, R. and Mills, E. (2012). Predictive value of cd4 cell count nadir on long term mortality in hiv positive patients in uganda, *HIV/AIDS Research and Palliative Care* .

Breslow, N. (1974). Covariance analysis of survival data under the proportional hazards model, *International Statistical Review* (43): 43–54.

Breslow, N. and J., C. (1974). A large sample study of the life table and product, *Collection of articles dedicated to Jerzy Neyman on his 80th birthday* pp. 437–453.

Cheng, R. (2012). Institutional characteristics and college student dropout risks: A multilevel event history analysis, *Res High Educ* **53**: 487–505.

Cox, D. R. (1972). Regression models and life-tables, *Journal of the Royal Statistical Society* **34**(2): 187–220.

Cox, D. R. (1975). Partial likelihood, *Biometrika* **62**(2): 269–276.

DesJardins, S. L., Ahlburg, D. A. and McCall, B. P. (1998). An event history model of student departure, *Economics of Education Review* pp. 375–390.

Efron, B. (1977). The efficiency of cox's likelihood function for censored data, *Journal of the American Statistical Association* (72): 557–565.

Greenwood, M. (1926). The natural duration of cancer, *Reports on Public Health and Medical Subjects* pp. 1–26.

Heagerty, P. J. (2005). Survival model predictive accuracy and ROC curves, *Biometrics* **61**: 92–105.

Heagerty, P. J., Lumley, T. and Pepe, M. (2000). Time-dependent ROC curves for censored survival data and a diagnostic marker, *Biometrics* **56**: 337–344.

Jensen, P. and Buddelmeyer, H. (2008). Innovation, technological conditions and new firm survival, *The Economic Record* **84**(267): 434–448.

Kaplan, E. L. and Meier, P. (1958). Nonparametric estimation from incomplete observations, *Journal of the American Statistical Association* **53**(282): 457–481.

McCall, B. P. (1994). Testing the proportional hazards assumption in the presence of unmeasured heterogeneity, *Journal of Applied Econometrics* **9**: 321–334.

Muthn, B. and Masyn, K. (2005). Discrete-time survival mixture analysis, *Journal of Educational and Behavioral Statistics* **30**(1): 27–58.

Peterson, A. V. (1977). Expressing the kaplan-meier estimator as a function of empirical subsurvival functions, *Journal of the American Statistical Association* **72**: 854–858.

Prentice, R. L. and Gloeckler, L. A. (1978). Regression analysis of grouped survival data with application to breast cancer data, *Biometrics* **34**: 57–67.

Radcliffe, P. M., Huesman, R. L. and Kellogg, J. P. (2006). Modeling the incidence and timing of student attrition: A survival analysis approach to retention analysis, *Association for Institutional Research in the Upper Midwest, Bloomington, Minnesota, USA* .

Scheike, T. H. and Jensen, T. K. (1997). A discrete survival model with random effects: An application to time to pregnancy, *Biometrics* **53**(1): 318–329.

Schoenfeld, D. (1980). Chi-squared goodness-of-fit tests for the proportional hazards regression model, *Biometrika* **67**(1): 145–153.

Singer, J. D. and Willett, J. B. (1993). It's about time - using discrete-time survival analysis to study duration and the timing of events, *Journal of Educational Statistics* **18(2)**: 155–195.

Smith, J. P. and Naylor, R. A. (2001). Dropping out of the university: a statistical analysis of the probability of withdrawal for uk university students, *Journal of the Royal Statistical Society* (164): 389–405.

Susarla, V. and Van Ryzin, J. (1976). Nonparametric bayesian estimation of survival curves from incomplete observations, *Journal of the American Statistical Association* **71**(356).

Therneau, T. M. and T., L. (2014). Package survival (R) (version 2.37-7), *R libraries* .

Vallejos, C. and Steel, M. (2014). Bayesian survival modelling of university outcomes.

Zhao, X. and Zhou, X. (2008). Discrete-time survival models with long-term survivors, *Statistics in Medicine* **27**: 1261–1281.