Open Journal of Statistics, 2016, 6, 908-916
http://www.scirp.org/journal/ojs
ISSN Online: 2161-7198
ISSN Print: 2161-718X

Study of University Dropout Reason Based on Survival Model

Juan C. Juajibioy

Fundación Universidad Autónoma de Colombia, Bogotá, Colombia
Email: juan.juajibioy@fuac.edu.co

How to cite this paper: Juajibioy, J.C. (2016) Study of University Dropout Reason Based on Survival Model. *Open Journal of Statistics*, **6**, 908-916.
http://dx.doi.org/10.4236/ojs.2016.65075

Received: July 28, 2016
Accepted: October 21, 2016
Published: October 24, 2016

Abstract

In this paper, we introduce the survival modelling methodology in order to identify some factors which may be influencing the university dropout. By using the data base provided by the Fundación Universidad Autónoma de Colombia and the semi parametric proportional hazard Cox model, we have been able to identify these risk factors.

Keywords

Dropout, Survival Models

1. Introduction

According to SPADIES[1] in Colombian Institutions Higher Education, around 20% of students beginning an undergraduate program drop out at first year. That is a global phenomenon: usually the group of graduates is smaller respect to the number of beginners. That is due to variables of academic, social or economic type and several studies have been realized about it. From this global phenomenon arose two big questions:

- What are the factors influencing the student drop out?
- How long take a student to drop out university?

The most literature about the first question is divided in two branches: Tinto's student integration model and Bean and Metzner's student attrition model (1985). The first one refers to the student's integration process and the second one refers to the student's individual variables, see [1] [2] and references therein for a detailed description.

Respect to the second question, the survival models have been amply developed, and typically focused on time to event data.

[1]Sistema para Prevención de la Deserción de la Educación Superior

2. Discrete Duration Analysis

Following [3] [4] we introduce the necessary background. Let T be the discrete variable representing the duration of studies (by semester from 1 until 12). The survival function is defined as

$$S(t) = P(T > t). \tag{1}$$

Since $p(t_k) = P(T = t_k)$ we have

$$S(t) = P(T > t) = \sum_{t_k > t} p(t). \tag{2}$$

The Hazard function is defined as

$$h(t_k) = P(T = t_k \mid T > t_{k-1}) = \frac{p(t_k)}{S(t_{k-1})}. \tag{3}$$

Notice that $P(T \geq t_k) = S(t_{k-1})$, since $p(t_k) = S(t_{k-1}) - S(t_k)$, by using (3) we have

$$\frac{S(t_k)}{S(t_{k-1})} = 1 - h(t_k), \tag{4}$$

so, the survival function can be written as

$$S(t) = \prod_{t_k \leq t} (1 - h(t)) \tag{5}$$

2.1. The Nonparametric Kaplan-Meyer Estimator

Let t_i the failure time, d_i the number of events that occur at time t_i and n_i the number of individuals at risk of experiencing the event immediately prior to t_j, then the product limit estimator of survival function is

$$\hat{S}(t) = \prod_{t_j < t} \left(\frac{n_j - d_j}{n_j} \right). \tag{6}$$

An interesting representation is given in [3] by using the following table

t_j	n_j	m_j	$\hat{S}(t_j)$
$t_0 = 0$	n_0	0	1
$\vdots$			
t_k	n_k	m_k	$\hat{S}(t_k)$

where n_0 is the initial population.

2.2. The Nonparametric Cox's Proportional Hazard Model

The Cox's proportional hazard model really gives a semi parametric method to the estimate the hazard function at time t given a baseline hazard that's modified by a set of covariates:

$$h(t \mid X) = h_0(t) \exp(\beta_1 X_1 + \cdots + \beta_n X_n) = h_0(t) \exp(\beta X) \tag{7}$$

where $h_0(t)$ is the non-parametric baseline hazard function $X = (X_1, \cdots, X_n)$ is a set of explanatory variables

3. Data and Descriptive Analysis

In this section we defined the principal explanatory variables and consider some descriptive aspects of these variables. We take a set that belong a cohort of students that began the studies in the first semester of 2010 in the University Fundación Universidad Autónoma de Colombia. In order to differentiate the group of students, we consider the following groups

- Group 1, Graduated Students: Student which finished successful their studies before 12 semesters.
- Group 2, Active students: In the dataset in second semester of 2015.
- Group 3, Inactive Students: Students who did not register for more than three consecutive semesters in the dataset.

In our analysis the following covariates were collected, grouped by individuals and academics. We consider the following individual variables

Variables		
Individuals	Gender	0 for female and 1 for male.
	Age	Age of the student when beginning his studies.
	Social status	In Colombia there are six class of social status.
	Location	Location of student's home.
academics	P1	Grade point average at first semester.
	P2	Grade point average at second semester.
	P3	Grade point average at second semester.
	Picfes	Score in icfes tests.

A breakdown by program and group is given in **Figure 1**. And in **Figure 2**, we show the percent of students by program.

In **Figure 2** we present the percent of students that began their studies at first semester of 2010.

The student population considered in this study, initially counted with 1018 students and due to the lack of information concerning to the explanatory variables we only considered a total population of 991 students. The total of students who dropped out in the period corresponding to first semester of 2010 until second semester of 2015 was of 37.54%, in **Figure 3** we show the distribution by groups. The Fundación Universidad Autónoma de Colombia is divided in four big faculties namely, Faculty of Law, Engineer Faculty, Faculty of Management and Accounting sciences and Human Science Faculty. In **Figure 1** (left square) can see that the bigger percent of students that dropped out university was in Law Faculty (8.6% in group 3).

4. Duration Analysis

In this section we looking for the relationship between the student's decision to complete or abandon, opposite to the decision of prolong their permanence at university.

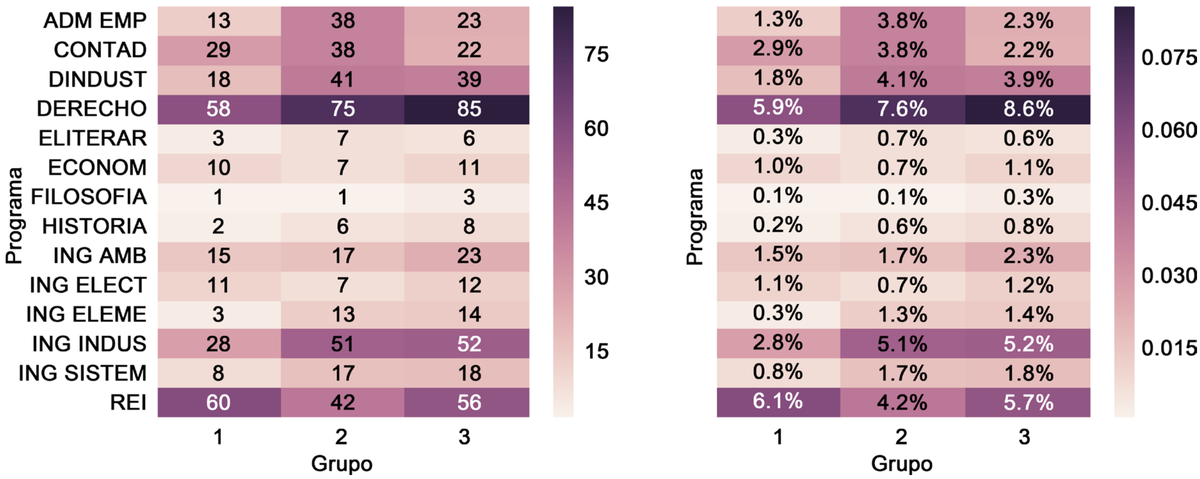

Figure 1. Breakdown by program and group.

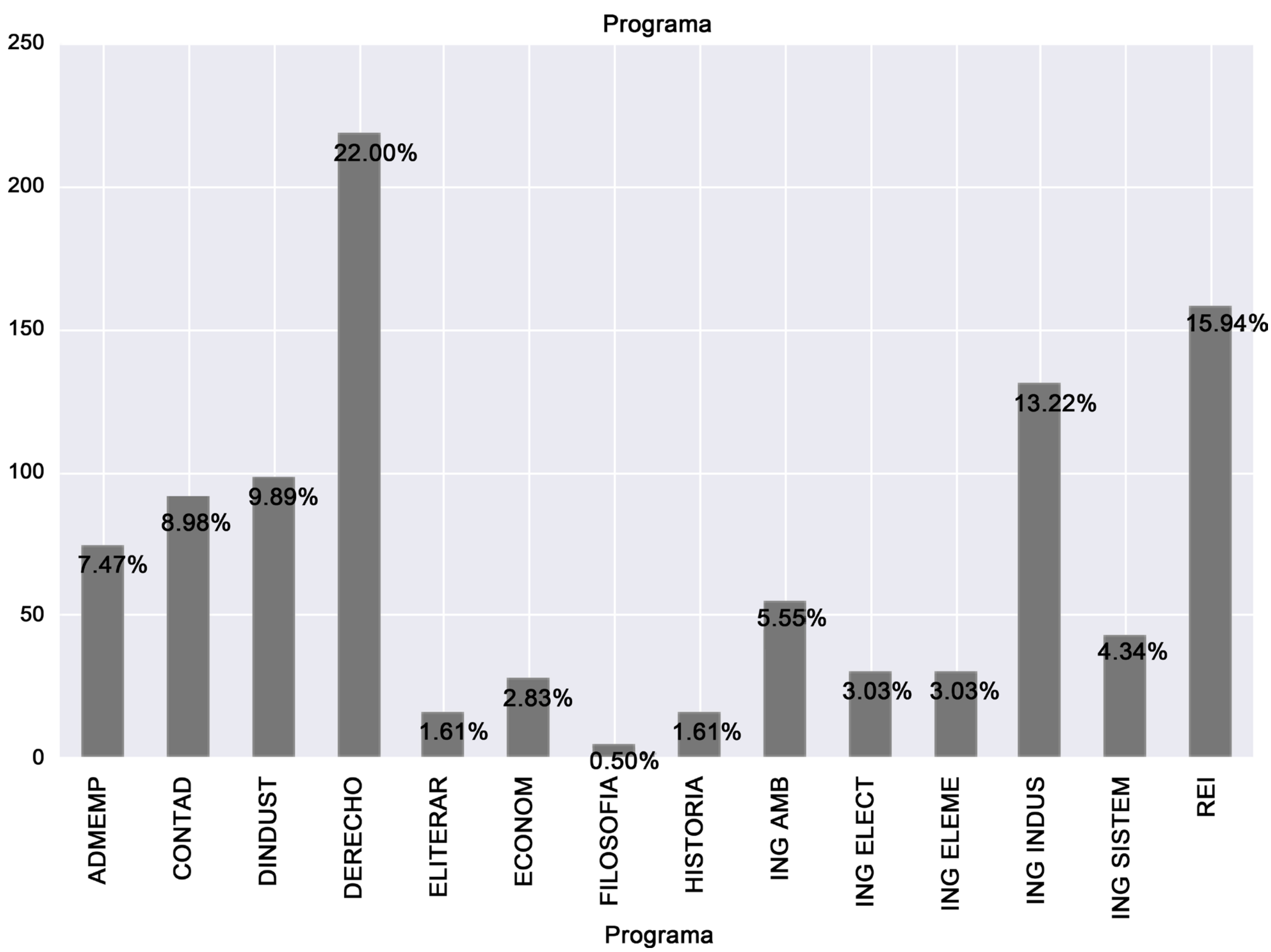

Figure 2. Distribution of students by program.

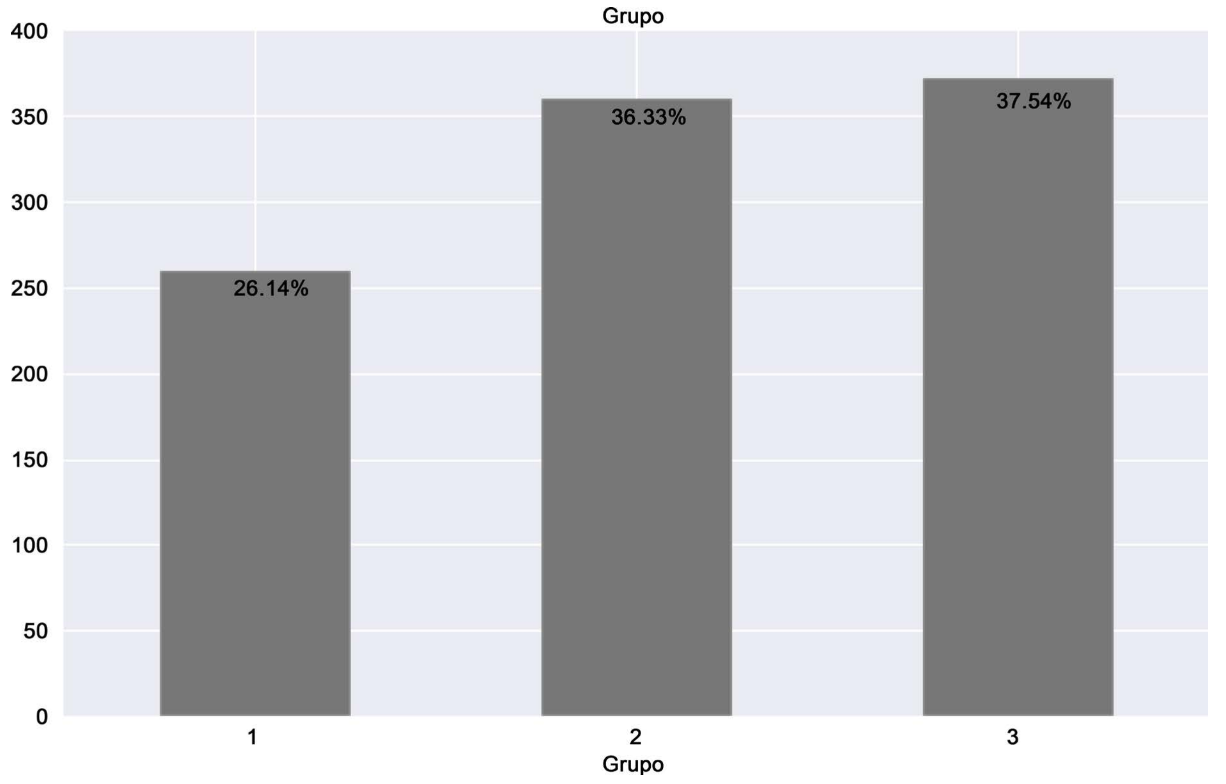

Figure 3. Distribution of students by group.

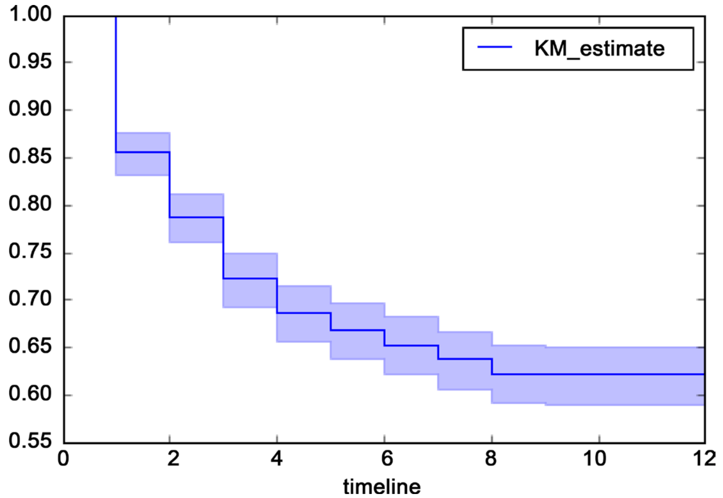

Figure 4. Kaplan Meier estimate for Survival function.

Initially we used the nonparametric Kaplan-Meier estimator 2.6, the results are given in Table 1 (See **Appendix**)

In Figure 4 it can see that the bigger drooping out rate occurs during the four initial semesters. In Figure 5 it is possible see the dynamics of survival in all programs that university offers

In order to study the effect of covariates we use the proportional hazard Cox model. In order to choice the significant variables we use the likelihood test ratio, the final

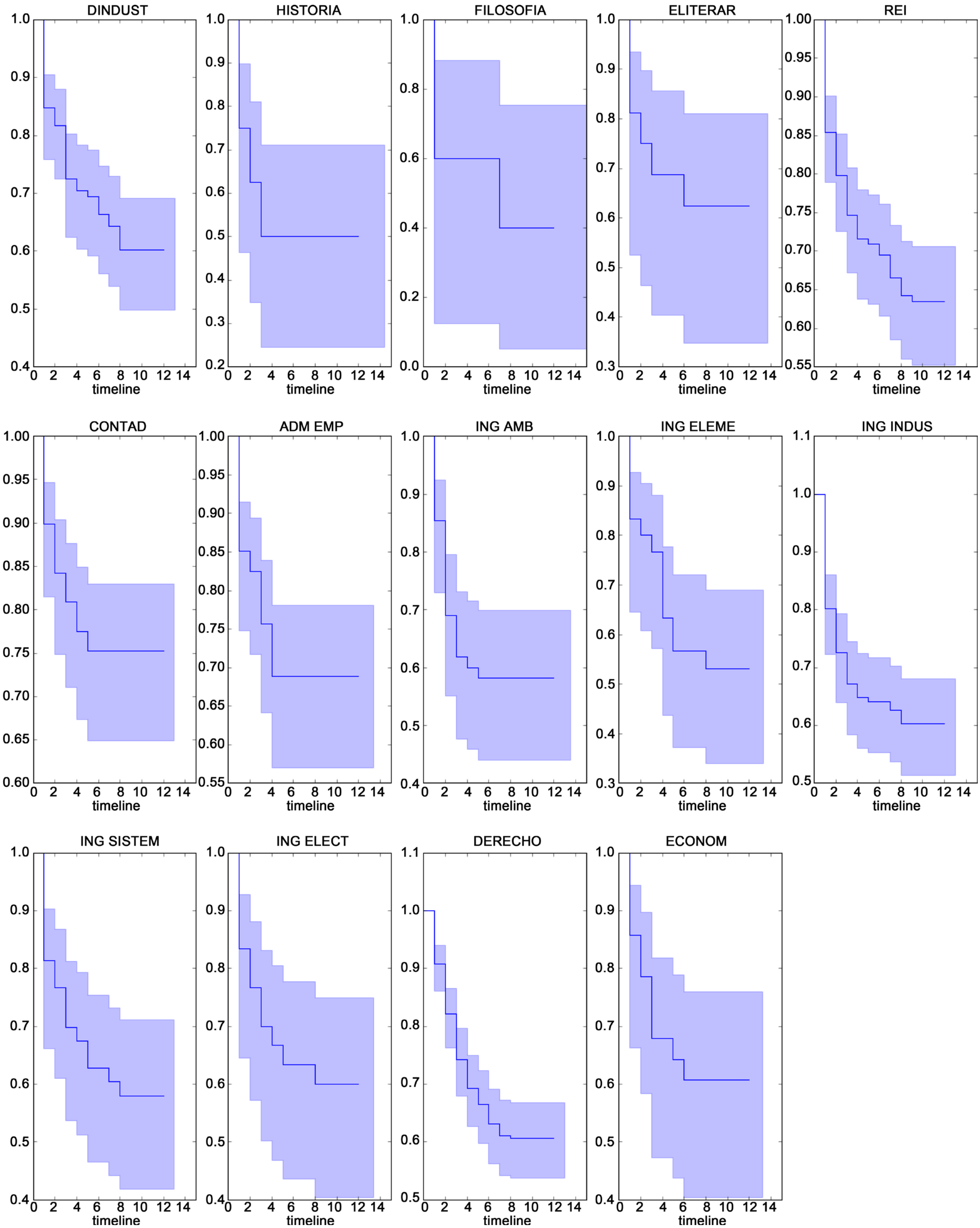

Figure 5. KM estimate by program.

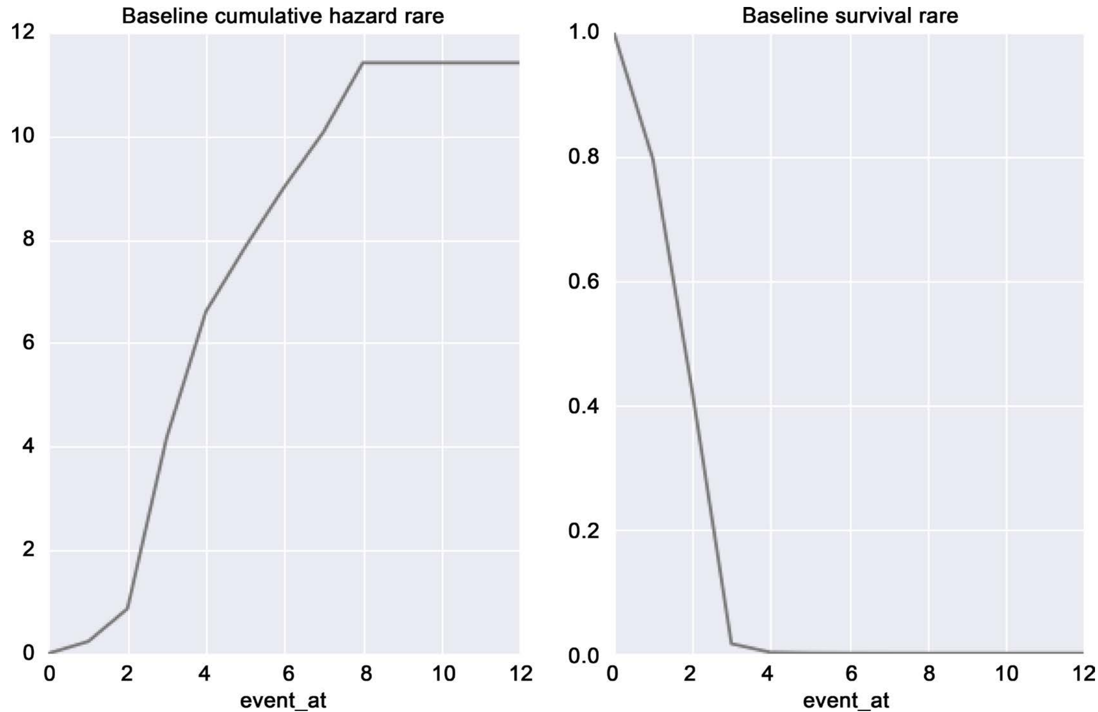

Figure 6. Baseline cumulative hazard and survival rate.

results can see in Table 2 (See **Appendix**)

The baseline cumulative hazard $H(t)=\sum_{t_j<t} h_0(t)$ it can see in Figure 6, notice in the left side the rapidly increasing rate, meaning that the hazard increase during the four first semesters.

5. Conclusion

In this work, we use the nonparametric survival model in order to estimate the risk factors for the university drop out, factors such that grade point average at first semester, gender and location are most significant in our study, remember that a positive estimate in the coefficient indicates an increased hazard meaning shorter expected survival time. By gender, the male population has more hazards to survival than female population. Finally after accounting for age, sex, grade point average and location there are no statistically significant associations between Icfes score and Social status and all-cause drop out.

Acknowledgements

This research was supported by SUI: Sistema Universitario de Investigación, Fundación Universidad Autónoma de Colombia.

Conflict of Interest

The authors declare that there is no conflict of interests regarding the publication of this paper.

References

[1] Montoya Diaz, M. (1999) Extended Stay at University: An Application of Multinomial Logit and Duration Models. *Applied Economics*, **31**, 1411-1422. http://dx.doi.org/10.1080/000368499323292

[2] Giovagnoli, P. (2005) Determinants in University Desertion and Graduation: An Application Using Duration Models. *Ecónomica LI*, No. 1, 60-90.

[3] Kleinbaum, D. and Klein, M. (2005) Survival Analysis: A Self-Learning Text. Springer.

[4] Pintilie, M. (2006) Competing Risks: A Practical Perspective. Wiley. http://dx.doi.org/10.1002/9780470870709

Appendix

Table 1. KM Estima for survival function.

t_j	$\hat{S}(t_j)$
0	1.000000
1	0.855701
2	0.788093
3	0.722503
4	0.686176
5	0.667850
6	0.653172
7	0.637957
8	0.622397
9	0.621255
10	0.621255
11	0.621255
12	0.621255

Table 2. Hazard ratios.

	coef	exp(coef)	se(coef)	z	p	lower 0.95	upper 0.95
BARRIOS UNIDOS	−0.946222	0.388205	1.098491	−0.861384	3.89E−01	−3.099698	1.207253
BOSA	−0.98285	0.374243	0.615371	−1.597167	1.10E−01	−2.189219	0.22352
CANDELARIA	0.539746	1.715571	0.585012	0.922625	3.56E−01	−0.607108	1.6866
CHAPINERO	0.855649	2.352901	0.641721	1.333366	1.82E−01	−0.402377	2.113675
CIUDAD BOLIVAR	−0.667607	0.512934	0.649726	−1.027521	3.04E−01	−1.941327	0.606113
ENGATIVA	0.349825	1.418819	0.486708	0.718757	4.72E−01	−0.604316	1.303965
FONTIBON	−0.616307	0.539935	0.674569	−0.91363	3.61E−01	−1.938729	0.706116
KENNEDY	−0.324605	0.722813	0.494109	−0.656951	5.11E−01	−1.293253	0.644043
LOS MARTIRES	−0.523431	0.592484	0.838874	−0.623968	5.33E−01	−2.167956	1.121094
PUENTE ARANDA	0.046525	1.047625	0.59174	0.078624	9.37E−01	−1.113519	1.20657
RAFAEL URIBE URIBE	−0.448711	0.63845	0.576947	−0.777734	4.37E−01	−1.579755	0.682332
SAN CRISTOBAL	0.042609	1.043529	0.528241	0.080661	9.36E−01	−0.992951	1.078169
SANTA FE	−0.818594	0.441051	0.735878	−1.112406	2.66E−01	−2.261205	0.624016
SOACHA	−0.481271	0.617997	0.741438	−0.649105	5.16E−01	−1.934783	0.972241
SUBA	0.409114	1.505484	0.51991	0.786895	4.31E−01	−0.610114	1.428343
TEUSAQUILLO	1.121985	3.070944	0.679139	1.652069	9.85E−02	−0.209396	2.453366
TUNJUELITO	−0.471024	0.624363	0.61123	−0.770616	4.41E−01	−1.669277	0.727229
USAQUEN	−0.151652	0.859287	0.573606	−0.264384	7.91E−01	−1.276147	0.972843
USME	−1.032805	0.356007	0.743826	−1.388504	1.65E−01	−2.490998	0.425387
P1	0.088902	1.092973	0.135613	0.655554	5.12E−01	−0.176953	0.354757
P2	−0.365178	0.694073	0.094174	−3.877699	1.05E−04	−0.549796	−0.18056
P3	−0.610764	0.542936	0.068857	−8.869989	7.32E−19	−0.745751	−0.475776
Picfes	−0.001673	0.998329	0.001826	−0.915817	3.60E−01	−0.005253	0.001908
Gender	0.198959	1.220132	0.164287	1.211043	2.26E−01	−0.123109	0.521027
Age	−0.018751	0.981424	0.018079	−1.037191	3.00E−01	−0.054192	0.01669
Social status	−0.357493	0.699427	0.098536	−3.628052	2.86E−04	−0.550662	−0.164324