Dropout and completion in the Bachelor in Statistics in Brazil: use of survival analyses in a competing risk scenario

Adriane Caroline Teixeira Portela[1], Andrea Diniz da Silva[2], Giovana Oliveira Silva[3]

Abstract: In the last decades, several public policies were created or reformulated in order to democratize access to higher education in Brazil. Although the positive impacts of these actions are visible, after entering higher education, students still face several challenges that affect their permanence at this educational level. It is necessary to identify which and how variables are associated with dropout and completion rates, because despite the abundant academic production on the subject, there are still gaps in the understanding of factors associated with specific courses, especially Statistics. Survival analysis were used to approach a situation in which dropout is competing with completion, in particular competing risk analyses was used. For this purpose, the cumulative incidence function was used, and the competitive risk model introduced by Fine and Gray (1999) was adjusted. Brazilian education census microdata was used and students who entered the year 2010 were followed up until 2017. The results found confirm the influence of relevant factors in the literature such as sex, age, shift, engagement in complementary activity and the type of academic organization.

keywords: dropout and completion; Bachelor in Statistics; survival analyses, competing risk analyses

1. Introduction

To democratize the access to higher education in Brazil, several public policies have been created or reformulated in recent years. Among them, we can mention the expansion of the offer to low-income, black and disable candidates in public higher education, as well as students' loans and other financial aid programs. Such intervention is urgent as the gross schooling rate in higher education in the country is one of the lowest in Latin America, although the privatization is one of the highest in the world (PINTO, 2004).

The reformulated and implemented programs have led to an increase in the Brazilian population in higher education, however it is estimated that only 17.4% of the population aged 25 or over had a complete degree in 2019 (PNAD, 2019). It happens because after entering higher education, students still face several challenges, which directly or indirectly affect their permanence or not at this educational level.

Dropout and completion in higher education has been a subject of major debates in recent decades, as it is of great interest to the public and private sectors. The factors associated with dropout and completion can be related to specific attributes of educational institutions and individual attributes of students, therefore, it is often necessary to study specific institutions and courses, to apply more effective public policies directed to the reality of these groups.

Considering the importance of the theme, this article seeks to outline the profile of students who are more likely to drop out and to complete a B. S. in Statistics in Brazil, as well as to identify which factors are associated with dropout and conclusion rates.

2. The Data Source and the Study Population

Higher Education Census (Censup) microdata were used. Censup is carried out annually by the National Institute of Educational Studies and Research Anísio Teixeira (Inep) in partnership with the Higher Education Institutions (HEI). The microdata analyzed refer to the students that started their Bachelor in Statistics in 2010 in one of the HEI in Brazil. These newcomers were followed up until

[1] Student at the National School of Statistical Sciences – ENCE/IBGE

[2] Professor at the National School of Statistical Sciences – ENCE/IBGE

[3] Professor at the Federal University of Bahia - UFBA

2017 only because from 2018 there was a recoding of the variable student code, making monitoring impossible in years following.

From the second term of the first year the students were classified as active, if they were currently enrolled or interrupted; as dropout if they have drop or enrolled another course in the same institution; and as graduate if they had completed their studies in the bachelor in Statistics.

In 2010, a total of 1,612 students started the bachelor in Statistics in one of the 32 HEI in Brazil. Among them, 991 (61.5%) are male and 621 (38.5%) are female. The average age was approximately 22.8 years (standard deviation of 6.7 years), with a minimum age of 16 years and a maximum of 65 years. The average number of students per HEI was 50.4. Attendance in evening count on 43% students, while 33% attended full-time, 22% in the morning and 2% in the evening. Over the total, 59 students participated in some type of complementary training activity.

3. Methods

Survival Analysis methods as presented by Carvalho et al. (2011) were considered, making dropout and conclusion as events of interest. In the literature, it is referred as competing risks. A function of great relevance in the survival analysis in the presence of competitive risks is the cumulative incidence function (CIF), also known as subdistribution, associated with the cause k, k = 1,…, K (F_k (t)), which provides at each time t the cumulative probability of outcome due to the k-th cause, considering the presence of all other causes. The CIF is defined by: $\hat{F}_k(t) = \sum_{j:t_j\leq t}\frac{d_{kj}}{n_j}\hat{S}(t_{j-1})$, where d_{kj} is the number of time faults tj associated with the cause k, n_j is is the number of individuals at risk in time tj and $S(t)$ is the general survival function.

Fine and Gray (1999) introduced a way to estimate the effect of covariates. They defined a subdistribution risk function as by directly modeling the CIF of a particular cause of event of interest k in an environment of competing risks, in which it is not necessary to assume independence between the times of competing events.: $\bar{\lambda}_k(t) = \lim_{\Delta t\to 0}\frac{P\{t \leq T<t+\Delta t, C=k|T\geq t \cup (T\leq t\cap C\neq k)\}}{\Delta t}$.

The variable C assumes zero if the individual's observation is censored, otherwise, C = k, where k indicates the first observed event of interest (k = 1, 2, ..., K). In addition, in the presence of a vector of covariates $\boldsymbol{x}$, the model assumes proportional risks, that is, the constant effect of covariates over time. The cumulative incidence function for cause k is given by: F_k $(t|\boldsymbol{x})$ = $P(T \leq t,K = k \mid \boldsymbol{x}) = 1 - \exp\{-\Lambda_{k,0}(t) \exp(\boldsymbol{x}^T\boldsymbol{\beta}_k)\}$, where T denotes the time variable, k indicates the event of interest, $\boldsymbol{x}$ = (x1, ..., xp) corresponds to a vector of covariates, $\boldsymbol{\beta}_k$ to a vector of regression coefficients associated with the cause k and Λk, 0 (t) a function of accumulated failure rate for specific cause k and not decreasing such that Λ_k, 0 (0) = 0.

Fine and Gray present a model under the assumption of proportional risks over subdistribution risks: $\bar{\lambda}_k(t|\boldsymbol{x}) = \bar{\lambda}_{k,0}(t)\exp\left(\boldsymbol{x}^T\boldsymbol{\beta}_k\right)$. The estimation of the parameter vector $\boldsymbol{\beta}$ is done by maximizing the partial likelihood function, which for $\bar{\lambda}_k(t|\boldsymbol{x})$ is given by: $\mathcal{L}(\boldsymbol{\beta}_k) = \prod_{k=1}^{K}\prod_{i=1}^{d_j}\frac{\exp\left(\boldsymbol{\beta}_k\boldsymbol{x}_{(k)i}^T\right)}{\sum_{l\epsilon R\left(t_{j(i)}\right)} w_{il}\exp\left(\boldsymbol{\beta}_j\boldsymbol{x}_l^T\right)}$.

The proposed method is similar to the usual one. The difference is considering as risk group individuals exposed to competing events, not just those free from any event. The weight w_{jl} changes every time tj in which the event of interest occurs, after all, the individual who has already experienced a competitive event cannot contribute in the same way as an individual who has not experienced any of the events.

Finally, for the correct use of the proposed Fine and Gray model, the assumption of proportionality of risks and linearity in the continuous variables that were considered in the final model was verified. The residuals of the adjusted models were visually analyzed and the correlation between the residuals and a function of time, in which it is expected not to verify patterns in the residual's graphs or the correlation with time, were checked as proposed by Pintilie (2006).

4. Results

To verify the characteristics of the students that are associated with dropout and conclusion the Cumulative Incidence Function (CIF) was used to show estimated probability of the outcomes. The computations were made using mstate and survival libraries of the R software (CARVALHO et al. 2011).

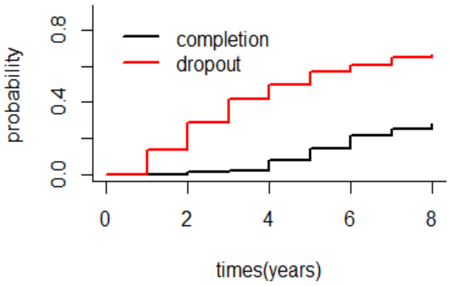

Figure 1 – Cumulative Incidence Function for completion and dropout

The curves in Figure 1 indicate that a student is more likely to experience dropout first than completion. In the first years, the completion curve (black) is expected to be close to zero, as the bachelor in Statistics lasts 4 years in Brazil, so students who graduate ahead of schedule, because of re-entry or transfer, are few. However, at the end of 8 years, the probability of the students having completed was approximately 0.27, while for the dropout was higher since the first years.

Six covariables were analyzed to see if there is a significant difference between the respective CIF, in order to profile the students who are most likely to experience dropout and completion. Figure 2 illustrates the comparison of the curves according to sex, to age being larger or smaller than the median age, to shift and to the condition of being engaged in complementary activities.

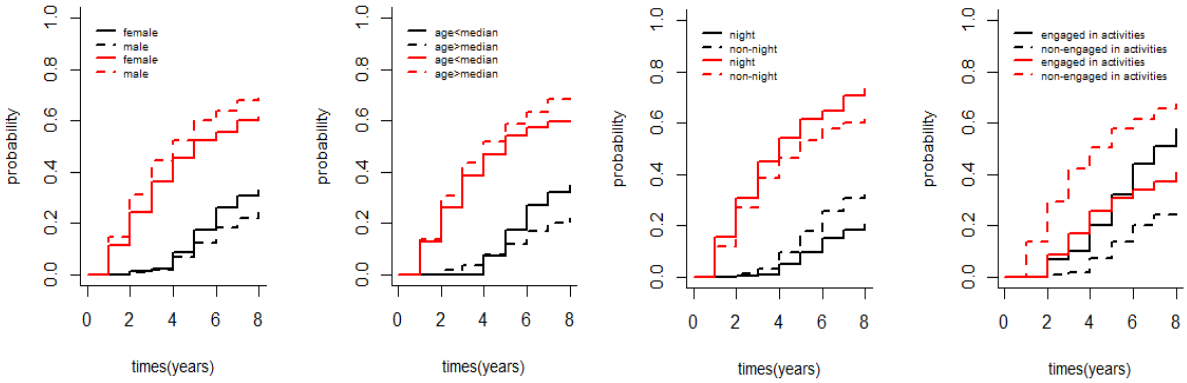

Figure 2 – Cumulative incidence functions for completion and dropout by covariables of interest

Curves differ from each other, at the significance level of 5%, what indicates it is possible to profile the students who are more likely to experience completion or dropout. Are more likely to complete the bachelor in Statistics (black lines) female, younger persons, non-night courses students and the ones engaged in complementary activities. Otherwise, male, older than the median age students, night courses students and non-engaged to complementary activities are more likely to dropout (red lines).

To assess the effects of covariables on completion or dropout rates, two models were adjusted, in which the first model the event of interest is the conclusion, and the competitive event is evasion. The second model the event of interest is the dropout and the competitive event is the conclusion. The linear correlation of the residues with time was tested, in which correlation values close to zero, show evidence in favor of the assumption of proportional risks, so not rejecting the null hypothesis (H0: $\rho = 0$) indicates that the covariate p meets the assumption. All tests indicated the proportionality of the risks ($p > 0.01$).

Table 1 - Estimates of the effects of covariables for the Fine-Gray model for conclusion and dropout

	Covariable	$\widehat{\beta}$	exp($\widehat{\beta}$)	z	p-value
Conclusion	Sex	0.369	1.446	3.845	< 0.01
	Age	-0.059	0.943	-5.493	< 0.01
	TypeHEI	-0.594	0.552	-3.796	< 0.01
	Shift	-0.470	0.625	-4.523	< 0.01
	Activity	1.023	2.782	5662	< 0.01
Dropout	Sex	-0.219	0.803	-3.423	< 0.01
	Age	0.019	1.018	4.588	< 0.01
	Shift	0.232	1.261	3.747	< 0.01
	Activity	-0.799	0.449	-3.865	< 0.01

Data Source: INEP. Censo da Educação Superior - Censup, 2010.

The completion rate for the bachelor in Statistics in Brazil for women is 1.45 times the completion rate for men and each one-year increase in age of the student results in a decrease of 6% to the completion rate. The student who is enrolled in a university-type HEI has a completion rate reduced 45% when compared to a student from an HEI whose academic organization is classified as a non-university (academic). For students who attend the evening course, the completion rate is 37% lower than those who attend during the day. Finally, the student being involved in complementary activities in the first year of graduation increases the completion rate by 2.78 times when compared to students who do not do complementary activities (Table 1).

For the second model, it is observed that the dropout rate of the bachelor in Statistics for women is 20% lower than the dropout rate for men and each one-year increase in age, the dropout rate increases by 1.8%. For students who attend the evening course the dropout rate is 1.26 times the dropout rate for students who attend day course and, finally, students who were engaged in complementary activities in the first year of graduation have the dropout rate 55% lower than students who were not.

5. Final Remarks

The present work had as proposal to identify variables that influence the completion and the dropout in the bachelor in Statistics in Brazil. The results are compatible to those found in the literature for other courses, despite the different techniques used. The study pointed out that the student's sex and age, the type of academic organization of the HEI, the shift attended, and the practice of complementary activity have significant effects on the completion. For dropout, it was found that sex, age, shift and complementary activity were the significant covariables.

An important highlight of this study is the strong influence of participation in complementary activities on the completion rate. The results discussed here show that participating in research, extension and teacher assistant activities dramatically increases the completion rate, so that reduces the dropout rate. Therefore, it reinforces the idea that investing in the offer of complementary activities can be seen as an alternative to fight dropout to be employed in the first years of the bachelor in Statistics.

ACKNOWLEDGEMENT

The first author thanks Universidade de São Paulo and Universidade Federal de São Carlos, for supporting her participation in the 63rd ISI World Statistics Congress.

References

Carvalho, M.S. et al. (2011). Análise de sobrevivência: teoria e aplicações em saúde. 2 ed. Rio de Janeiro. Editora Fiocruz. 432p.

Fine, J. P.; Gray, R. J. (1999). A proportional hazards model for the subdistribution of a competing risk. Journal of the American Statistical Association, New York, v. 94, n. 446, p. 496-509.

Pesquisa Nacional por Amostra de Domicílios (PNAD) 2019. Instituto Brasileiro de Geografia e Estatística – IBGE. Brasil.

Pintilie, M. (2006). Competing Risks: a practical perspective. Chichester: John Wiley & Sons.

Pinto, José Marcelino de Rezende. O acesso à educação superior no Brasil. Educ. Soc., Campinas, v. 25, n. 88, p. 727-756, Out. 2004.