https://eventos.utfpr.edu.br//sicite/sicite2020

Análise de sobrevivência aplicada a dados de evasão dos alunos em um curso de Licenciatura em Matemática

Survival analysis applied to student dropout data in a Mathematics Degree course

Kimberly de Azevedo
kimberlyazevedo@alunos.utfpr.edu.br
Universidade Tecnológica Federal do Paraná, Cornélio Procópio, Paraná, Brasil.

Roberto Molina de Souza
rmolinasouza@utfpr.edu.br
Universidade Tecnológica Federal do Paraná, Cornélio Procópio, Paraná, Brasil.

Recebido: 19 ago. 2020.
Aprovado: 01 out. 2020.

RESUMO

Análise de sobrevivência é um método estatístico utilizado em várias áreas do conhecimento. É considerada a metodologia estatística mais adequada para analisar dados temporais até a ocorrência de um evento de interesse. O objetivo deste trabalho foi estudar o tempo até um estudante evadir no curso de Licenciatura em Matemática da Universidade Tecnológica Federal do Paraná, Campus Cornélio Procópio (UTFPR-CP). Para descrever os dados foram usados os estimadores de Kaplan-Meier e para analisá-los a modelagem paramétrica com as distribuições exponenciais e Weibull. As análises foram realizadas utilizando o software livre R. Observou-se que os estudantes casados ou com união estável ou que não participam de algum projeto têm maior chance de evadir ao longo de cada semestre quando comparados aos estudantes solteiros ou divorciados ou que participam de projetos, respectivamente. Este trabalho é de interesse para a coordenação do curso de Licenciatura em Matemática da UTFPR-CP no sentido de observar com mais cuidado os estudantes que podem vir a evadir em um tempo menor e tomar medidas necessárias para evitar a evasão.

PALAVRAS-CHAVE: Evasão universitária. Software livre. Estatística matemática.

ABSTRACT

Survival analysis is a statistical method used in several areas of knowledge. It is considered the most appropriate statistical methodology for analyzing time data until the occurrence of an event of interest. The objective of this work was to study the time until a student escapes in the Mathematics Degree course at the Federal Technological University of Paraná, Câmpus Cornélio Procópio, (UTFPR-CP). To describe the data, the Kaplan-Meier methods and parametric modeling with exponential and Weibull distributions were used to study the time until evasion. The analyzes were performed using the free software R. It was observed that students who are married or in a stable relationship or who do not participate in any project are more likely to drop out during each semester when compared to single or divorced students or who participate in projects, respectively. This work is of interest to the institution in coordinating the Mathematics Degree course at UTFPR-CP in order to observe more carefully the students who may evade in less time.

KEYWORDS: University dropout. Free software. Mathematical statistics.

INTRODUÇÃO

Análise de sobrevivência ou *Survival analysis* é um método estatístico muito utilizado em ciências da saúde, biológicas e engenharias, possuindo aplicações em outras áreas do conhecimento. É considerada a metodologia estatística mais adequada para lidar com dados temporais até a ocorrência de um evento de interesse (tempo de falha), na presença de censuras, as quais são consideradas observações incompletas e se caracterizam como dados de indivíduos em que a ocorrência do evento não foi verificada (COLOSIMO; GIOLO, 2006). A censura é o fator que diferencia a análise de sobrevivência de outras técnicas de análise de dados.

O objetivo deste trabalho é estudar a análise de sobrevivência e confiabilidade bem como a aplicação deste ferramental estatístico em problemas aplicados fazendo o uso do software livre R. Os objetivos específicos foram: estudar a análise de sobrevivência e confiabilidade bem como seus aspectos computacionais (software R); obter um conjunto de dados para modelagem estatística e aplicação das ferramentas, e analisar os dados, comparar os modelos e apresentar as conclusões.

METODOLOGIA

Na análise de sobrevivência, a estimativa da função de sobrevivência de um evento de interesse é comumente obtida utilizando o método de Kaplan e Meier (1958), e a influência das covariáveis sobre o tempo de sobrevivência pode ser examinada através do modelo semi-paramétrico de Cox (1972) sobre a função de risco da causa específica ou ainda sobre modelos exclusivamente paramétricos.

Também conhecido como estimador limite-produto, o estimador Kaplan e Meier é considerado não paramétrico (independente de uma função distribuição de probabilidade) para a função de sobrevivência. Baseado apenas nas informações dos dados, sendo assim um estimador empírico, na ausência de censuras é definido por:

$$\hat{R}(t) = \frac{\#r_t}{n} \tag{1}$$

em que $\#r_t$ é o número de observações que ainda não sofreram o evento de interesse até o tempo t; e n o número total de observações em estudo. Este estimador tem a forma de uma escada.

Considerando tempos censurados, ou seja, quando algumas unidades amostrais não experimentam o evento de interesse, o gráfico de Kaplan e Meier, em sua forma de escada, tem os valores censurados representados por um sinal de "+" onde ocorrem no gráfico. A construção destes gráficos é bastante facilitada utilizando a função *survfit* do pacote *survival* do software livre R (CORE TEAM, 2020).

Como o estimador de Kaplan e Meier é uma ferramenta gráfica, é comum o uso do teste *Logrank*, proposto por Mantel (1966), que compara as estimativas das

funções de risco de dois grupos (ou níveis) de uma covariável em cada tempo do evento observado, a um dado nível de significância.

Considere uma covariável com dois níveis. Seja $1,2,\dots,J$ os diferentes tempos do evento de interesse observado em cada grupo. Seja, $N_{1,j}$ e $N_{2,j}$ o número de sujeitos em risco (que ainda não sofreram o evento de interesse ou que foram censurados) no início do j −ésimo período, em cada grupo, respectivamente. Seja $O_{1,j}$ e $O_{2,j}$ o número observado de eventos nos grupos, no tempo t. Logo, $N_j = N_{1,j} + N_{2,j}$ e $O_j = O_{1,j} + O_{2,j}$.

O teste *Logrank* tem como hipótese nula (H_0) que o risco entre os grupos são os mesmos. Caso a hipótese nula seja rejeitada ao nível α* de significância previamente fixada, assume-se que existem evidências que a covariável em questão é relevante para o evento de interesse.

Da forma com que está representado, $O_{i,j}$ segue distribuição hipergeométrica com parâmetros N_j, $N_{i,j}$ e O_j com:

$$E_{i,j} = N_{i,j}\frac{O_j}{N_j} \tag{2}$$

$$V_{i,j} = E_{i,j}\left(\frac{N_j - O_j}{N_j}\right)\left(\frac{N_j - N_{i,j}}{N_j - 1}\right)$$

Logo, a estatística do teste *Logrank* como uma aproximação da distribuição normal, para todo $j = 1,2,\dots,J$, $i = 1,2$, sob a hipótese nula H_0 é dada por:

$$Z_c = \frac{\sum_{j=1}^{J}(O_{i,j} - E_{i,j})}{\sqrt{\sum_{j=1}^{J} V_{i,j}}} \sim N(0,1) \tag{3}$$

em que se rejeita H_0 se $P(|Z_c|) \geq \alpha^*$.

A construção dos gráficos de Kaplan e Meier pelos níveis da covariável juntamente com o teste *logrank* é facilitada utilizando a função *ggsurvplot* do pacote *survminer* do software livre R.

Para tratar os dados de forma paramétrica, consideram-se distribuições de probabilidade. A seguir, serão apresentadas as distribuições exponencial e Weibull.

Seja T uma variável aleatória contínua que representa o tempo até a ocorrência de um evento. T segue distribuição exponencial com parâmetro α, $T \sim Exp(\alpha)$, se:

$$f(t) = \frac{1}{\alpha} exp\left(-\frac{t}{\alpha}\right); t \geq 0 \tag{4}$$

A função de sobrevivência da distribuição exponencial é dada por:

$$S(t) = exp\left(-\frac{t}{\alpha}\right) \tag{5}$$

A função de risco da distribuição exponencial é dada por:

$$h(t) = \frac{1}{\alpha} \tag{6}$$

Seja T uma variável aleatória contínua que representa o tempo até a ocorrência de um evento. T segue distribuição de Weibull com parâmetros α e β, $T \sim Wei(\alpha, \beta)$, se:

$$f(t) = \frac{\beta}{\alpha}\left(\frac{t}{\alpha}\right)^{\beta-1} exp\left(-\frac{t}{\alpha}\right)^{\beta}; t \geq 0 \quad (7)$$

em que α é o parâmetro de escala e β é o parâmetro de forma; $\alpha > 0$ e $\beta > 0$. Note que, $Wei(\alpha, 1) = Exp(\alpha)$.

A função de sobrevivência da distribuição de Weibull é dada por:

$$S(t) = exp\left(-\frac{t}{\alpha}\right)^{\beta} \quad (8)$$

A função de risco da distribuição de Weibull é dada por:

$$h(t) = \frac{\beta}{\alpha}\left(\frac{t}{\alpha}\right)^{\beta-1} \quad (9)$$

Em muitas aplicações em análise de sobrevivência o interesse vai além de considerar apenas o efeito de uma variável com relação ao fenômeno de estudo, mas de várias variáveis e, neste sentido, modelos de regressão podem ser incorporados nas distribuições introduzidas até aqui. Seja:

$$log(T_i) = \gamma_0 + \gamma_1 x_{i1} + \cdots + \gamma_p x_{ip} + \beta\varepsilon_i \quad (10)$$

em que T representa a variável aleatória tempo até a ocorrência do evento em estudo; $X_1, \ldots, X_p$ representam as covariáveis regressoras do modelo; $\gamma_1, \ldots, \gamma_p$ o efeito destas covariáveis; ε_i é o erro aleatório com distribuição valor extremo (VE) com função densidade de probabilidade $f(\varepsilon) = exp[\varepsilon - exp(\varepsilon)]$ para $\varepsilon \in R$. Assim, $T \sim Wei(\alpha_X, \beta)$ em que $\alpha_X = exp(\gamma_0 + \gamma_1 x_{i1} + \cdots + \gamma_p x_{ip})$ e $log(T) \sim VE(\mu_x, \beta)$ em que $\mu_x = \gamma_0 + \gamma_1 x_{i1} + \cdots + \gamma_p x_{ip}$.

A estimação dos parâmetros pode ser vista com detalhes em Lawless (2011). Na aplicação deste trabalho foi utilizada a função *survreg* do pacote *survival* do software livre R. As hipóteses de interesse, ou seja, quais variáveis influenciam no tempo de até a ocorrência do evento de interesse são da forma:

$$\begin{cases} H_0 : \gamma_j = 0 \\ H_1 : \gamma_j \neq 0 \end{cases} \text{ para } j = 1, \ldots, k \quad (11)$$

Como os estimadores $\hat{\gamma}_1, \ldots, \hat{\gamma}_p$ tem distribuições aproximadamente normal por propriedades assintóticas, a hipótese nula será rejeitada se o valor $p \leq \alpha^*$. $p = 2 \times P(Z \leq -|z_c|)$, em que $z_c = \frac{\hat{\gamma}_j}{\sqrt{\frac{Var(\hat{\gamma}_j)}{n}}} \sim N(0,1)$; $\sqrt{\frac{Var(\hat{\gamma}_j)}{n}}$ é o erro padrão de $\hat{\gamma}_j$; $Var(\hat{\gamma}_j)$ é a variância do estimador $\hat{\gamma}_j$; n é o tamanho amostral; α^* é o nível de significância previamente fixado e $N(0,1)$ indica que z_c segue distribuição normal padrão.

APLICAÇÃO

Para aplicação da metodologia proposta foram utilizados dados referentes a evasão no curso de Licenciatura em Matemática da UTFPR-CP. Primeiramente submeteu-se um projeto ao Comitê de Ética em Pesquisa (CEP) com seres humanos

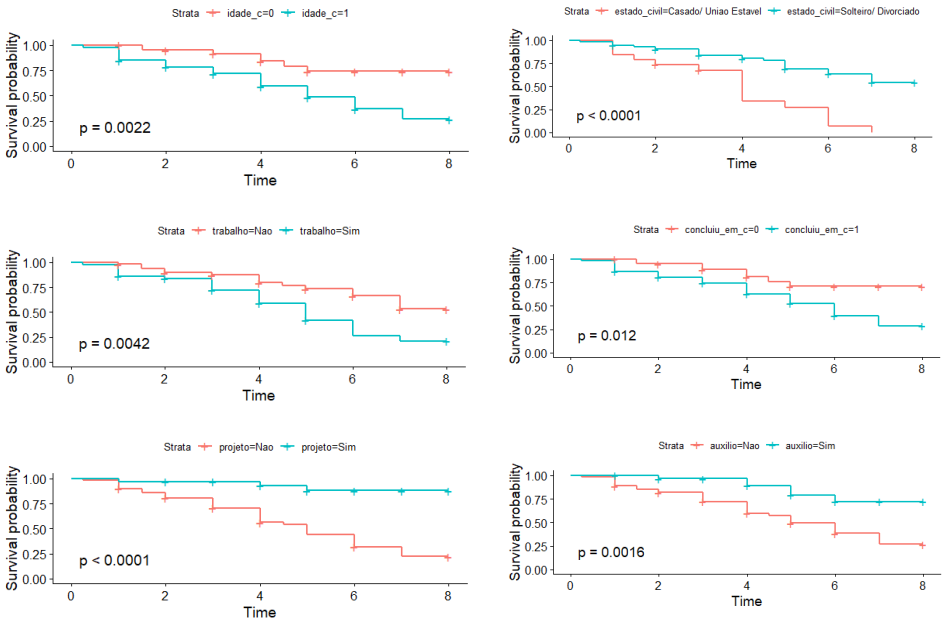
utilizando a plataforma Brasil conforme as resoluções 466, de 12 de dezembro de 2012 e 210 de 07 de abril de 2016. O projeto foi aprovado com o número CAAE 88428818.9.0000.5547. Após a aprovação pelo CEP foram aplicados dois questionários distintos: um para os estudantes que evadiram e outro para os estudantes que não evadiram. As variáveis analisadas neste trabalho foram comuns em ambos os questionários, porém houve questões específicas em cada um dos questionários que não foram tratadas neste trabalho.

Foi utilizado um estudo do tipo caso-controle, com uma amostra final constituída de $n = 93$ estudantes, sendo 38 estudantes evadidos e 55 estudantes não evadidos. Nesta aplicação, considera-se tempo completo, o semestre em que o estudante evadiu, sendo 38 observações e, dados censurados, o semestre em que o estudante está matriculado e não evadiu, sendo 55 observações.

ANÁLISES INICIAIS

Considerando o número de semestres como variável resposta, sendo os estudantes evadidos tempo completos e não evadidos tempo censurado e estratificando cada covariável em estudo em dois níveis, foram construídos os gráficos de Kaplan e Meier e calculado o valor p para a comparação dos riscos segundo os níveis de cada grupo utilizando o teste *logrank*, conforme pode ser observado na Figura 1, onde estão apresentadas apenas as variáveis significativas ao nível de 0,05 de significância.

Figura 1: - Gráficos de Kaplan e Meier e teste de logrank.



Fonte: Autoria própria (2020).

Nota: X_1=0 se a idade for menor que 24 anos ou X_1=1 se a idade for maior ou igual a 24 anos; X_2=0 se o estado civil é solteiro ou divorciado ou X_2=1 se o estado civil é casado ou em união estável; X_3=0 se não trabalha ou X_3=1 se trabalha; X_4=0 se concluiu o ensino médio há menos de 6 anos ou X_4=1 se concluiu o ensino médio há mais de 6 anos; X_5=0 se participa de projetos ou X_5=1 se não participa de algum projeto; X_6=0 se não recebe qualquer auxílio da universidade; X_6=1 se recebe algum auxílio da universidade.

MODELAGEM

Considerando apenas as variáveis relevantes do ponto de vista univariado, apresentadas na Figura 1, modelo de regressão proposto é dado por:

$$\alpha_X = exp(\gamma_0 + \gamma_1 X_{i1} + \gamma_2 X_{i2} + \gamma_3 X_{i3} + \gamma_4 X_{i4} + \gamma_5 X_{i5} + \gamma_6 X_{i6}) \quad (12)$$

e os parâmetros estimados utilizando o software livre R, com seus respectivos erros-padrão (E. P.), valor p e *Akaike Information Criterion* (AIC) (Akaike, 1974) são dados na Tabela 1 para a distribuição Exponencial e de Weibull.

Tabela 1 – Estimativas (Est.) dos parâmetros, erro-padrão (E. P.), valor *p* e AIC para a distribuição (dist.) Exponencial e de Weibull.

Parâmetros	Dist. Exponencial		Dist. de Weibull	
	Est. (E. P.)	Valor p	Est. (E. P.)	Valor p
γ_0	3,948 (0,675)	< 0,01	3,011 (0,458)	< 0,01
γ_1	-1,052 (0,825)	0,20	-0,633 (0,516)	0,22
γ_2	**-0,869 (0,394)**	**0,03**	**-0,577 (0,255)**	**0,02**
γ_3	-0,244 (0,388)	0,53	-0,185 (0,242)	0,44
γ_4	0,579 (0,791)	0,46	0,469 (0,494)	0,34
γ_5	**-1,395 (0,631)**	**0,03**	**-0,892 (0,394)**	**0,02**
γ_6	0,672 (0,465)	0,15	0,421 (0,293)	0,15
β	**1**	**-**	**0,605 (0,139)**	**<0,01**
AIC	219,4		212,6	

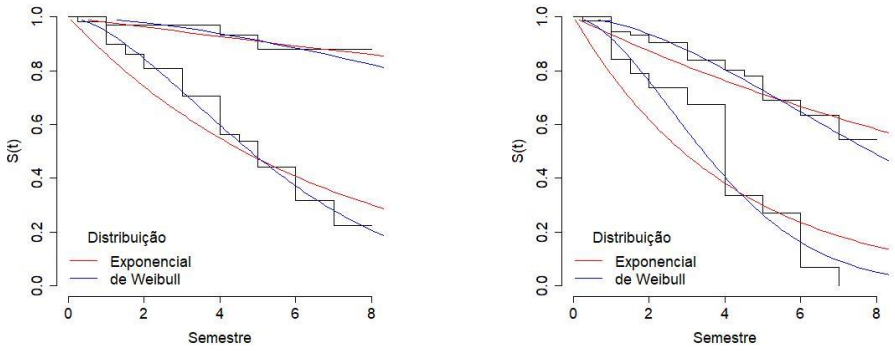
Fonte: Autoria própria (2020).

Na Tabela 1, observa-se que existem dois indícios para a escolha do modelo de Weibull: i) O valor de AIC da distribuição de Weibull $(212{,}6)$ é menor que o valor do AIC da distribuição exponencial $(219{,}4)$, indicando um melhor ajuste do modelo de Weibull; ii) Quando testada a hipótese nula $H_0 \ : \ \beta = 1$, a mesma é rejeitada $(p < 0{,}01)$, indicando que o parâmetro de forma da distribuição de Weibull é importante na modelagem.

Na Figura 2, pode-se observar também que o modelo de Weibull é mais bem ajustado às curvas de sobrevivência do que o modelo exponencial. Portanto, selecionado o modelo de Weibull, assumindo um nível $\alpha^* = 0{,}05$, as variáveis que apresentaram relevantes neste modelo foram estado civil e a participação do estudante em projetos.

Finalmente, em termos de interpretação dos parâmetros da Tabela 1, a atenção está naqueles correspondentes às variáveis que apresentaram evidência de significância, neste caso γ_2 e γ_5. Para que eles sejam interpretados na escala dos dados, faz-se $exp(\,0{,}577)^{0{,}605} = 1{,}42$ e $exp(\,0{,}394)^{0{,}605} = 1{,}27$ e pode-se concluir que: i) para cada semestre que se passa, o risco de um estudante casado ou em união estável evadir aumenta $1{,}42$ vezes quando comparados aos estudantes solteiros ou divorciados; ii) para cada semestre que se passa, o risco de um estudante que não participa de projetos evadir aumenta $1{,}27$ vezes quando comparado a um estudante que participa de projetos.

Figura 2: - Gráficos de sobrevivência com curvas ajustadas.



Fonte: Autoria própria (2020).

CONCLUSÕES

Como visto ao longo deste trabalho, principalmente no referencial teórico, o uso de análise de sobrevivência ou confiabilidade é importante quando a variável aleatória em estudo refere-se ao tempo até a ocorrência de um evento de interesse e, mais importante ainda quando incorporam-se neste estudo tempos incompletos, ou seja, levar em conta a informação daquelas unidades amostrais que ainda não experimentaram o evento de interesse.

Na aplicação, o uso das distribuições de Weibull e exponencial na modelagem do problema foram importantes para explicar questões relacionadas ao tempo de evasão dos alunos do curso de Licenciatura em Matemática da UTFPR-CP. Prevalecendo a distribuição de Weibull sobre a distribuição exponencial, pode-se concluir que o estado civil do estudante e a participação em projetos são relevantes na evasão, sendo que os estudantes casados ou em união estável ou que não participam de projetos têm maior chance de evadir ao longo de cada semestre quando comparados aos estudantes solteiros ou divorciados ou participam de projetos, respectivamente.

Finalmente este trabalho é de grande interesse para os gestores da instituição em medidas que visem diminuir a evasão dos estudantes do curso de Licenciatura em Matemática da UTFPR-CP, visto que proporcionar a participar de alunos em projetos em algo bastante factível, trazendo ao aluno a sensação de pertencimento. Por outro lado, dar uma atenção especial no acompanhamento dos alunos casados ou em união estável pode também diminuir a evasão deles.

AGRADECIMENTOS

Os autores agradecem a Fundação Araucária pelo fomento na forma de Bolsa de Iniciação Científica para a primeira autora deste trabalho. Agradecimento também pelo maravilhoso trabalho do orientador Prof. Dr. Roberto Molina, pela paciência e dedicação que teve com a orientanda.

REFERÊNCIAS

AKAIKE, H. A new look at the statistical model identification. **IEEE Transactions on Automatic Control**, v. 19, n. 6, p. 716-723, 1974. http://bayes.acs.unt.edu:8083/BayesContent/class/Jon/MiscDocs/Akaike_1974.pdf. Acesso em: 31 ago. 2020.

COLOSIMO, E.; GIOLO, S. **Análise de sobrevivência aplicada**. 1. ed. São Paulo: Edgar Blucher, 2006. (ABE – Projeto Fisher)

COX, D. R. Regression models and life-tables. **Journal of the Royal Statistical Society. Series B (Methodological)**, v. 34, n. 2, p. 187-220, 1972. http://www.biecek.pl/statystykaMedyczna/cox.pdf. Acesso em: 31 ago. 2020.

KAPLAN, E. L.; MEIER, P. Nonparametric estimation from incomplete observation. **Journal of the American statistical association**, v. 53, n. 282, p. 457-481, 1958.

https://web.stanford.edu/~lutian/coursepdf/KMpaper.pdf. Acesso em: 31 ago. 2020.

LAWLESS, J.; **Statistical Models and Methods for Lifetime Data**. 2. ed. New York: Wiley, 2011. (Wiley Series in Probability and Statistics)

MANTEL, N. Evaluation of survival data and two new rank order statistics arising in its consideration. **Cancer Chemotherapy Reports**, v. 50, n. 3, p. 163-170, 1966.

R Core Team. R: **A language and environment for statistical computing**. R Foundation for Statistical Computing, Vienna, Austria, 2020. https://www.R-project.org/. Acesso em: 25 ago. 2020.