Proceedings of the 7th International Conference on Statistics: Theory and Applications (ICSTA'25)
Paris, France - August, 2025
Paper No. 125
DOI: 10.11159/icsta25.125

A Survival Analysis of Adolescent Dropout in Swimming

Austin Yang
Charlotte Country Day School, Charlotte, NC, USA
ayang26@charlottecountryday.org

Abstract - *The benefits of adolescents participating in sports have been well recognized and documented for a long time. However, a significant concern in youth sports is the high dropout rate among young athletes. Anecdotal evidence suggests a lack of improvement over an extended period of time is one of the main factors that cause swimmers to leave the sport. This study is the first one that adopts a survival analysis framework to formally test these hypotheses. Using a large, publicly available database on competitive swimmers, this research examines how swimmers' performance improvement affects their decisions to quit or not. Analyzing nearly 12,000 swimmers' meet performances over the last 10 years, we create two metrics to track swimmer performance improvement. One measures swimmers' self-improvement, the other measures their relative improvement compared with peers. The main findings include (1) swimmers' absolute performance level and the speed of improvement both influence their dropout probability, with the absolute performance level being a more important factor; (2) swimmers who are faster when they are younger but slower when they grow up are more likely to quit; (3) if swimmers are "slow" relatively to their peers, regardless of how much progress they have made, they are more likely to quit. The main contribution of this project is using a large-scale database to empirically study swimmers' dropout behaviors as well as applying a survival analysis method to dropout in sports.*

Keywords: survival analysis; adolescent; sports dropout; large data; Cox Proportional Hazard model

1 Introduction

Sports are the most popular extracurricular activity for adolescents. The benefits of participating in sports on youth development, physically, mentally and socially, have been widely acknowledged and well documented [1]. However, adolescent dropout from sports has been an undeniable and long-standing issue [2]. This research aims to deepen the understanding of the factors affecting sports dropout, using large data.

A lot of research has been conducted to understand the dropout behaviors from social and psychological perspectives. However, the existing research is missing two important components. First, the importance of swimmers' progress, both self-improvements and improvements relative to peers, was not fully studied. This is because there is a lack of a good way to quantify youth athletes' progress and the existing research was not able to incorporate adolescents' long-time progress into the analysis. Second, research based on large data is scarce with most evidence-based research relies on survey or questionnaires [3-4]. Those samples usually consist of respondents from the same team or the same city, and the magnitude of those sample sizes is of a few hundred at most. The lack of large-scale data analysis significantly limits the generalization of those research findings.

Using the survival analysis method to analyze nearly 12,000 swimmers' meet performances over the last 10 years, our research finds that besides age and gender, swimmers' self-improvement and relative position among peers both have significant effect on their decisions to quit.

Our work contributes to the existing literature in the following ways. (1) It confirms and identifies that at certain ages, adolescents are more likely to quit than at other ages. (2) It confirms that female swimmers are more likely to quit than male swimmers. (3) It introduces a method to directly quantify swimmers' performance and demonstrates that both self-improvement and relative position changes among peers can influence their decisions to quit. (4) It finds that the absolute performance level is more important than improvement itself to explain dropout. (5) Under a survival analysis framework, this work establishes a case where swimmers' biological ages are more meaningful than "treatment" ages to explain dropout. The research expands the applicability of the survival analysis methodology.

1.1 Literature Review

While the reasons for dropout involve many aspects, researchers have mostly studied this subject from social and psychological perspectives. Some swimmers quit swimming because they switch to other activities [5]. Most reported direct reasons for quitting swimming are a lack of fun and less enjoyment. One factor causing this is training volume. Extensive training or hard training sets may cause swimmers to burn out. Research shows that swimmers enjoy practice more when there is more time off or play time [6]. In addition, lack of social support can make the sport less attractive. Social support could be verbal encouragement or expressions of caring from teammates and coaches. One-on-one coaching and team activities can all increase swimmers' feeling of social support [7]. Another factor causing dropout is pressure, possibly from competition. Concerns that they won't be able to perform well in swim meets can generate higher stress [4]. Pressure can also come from parents with some having been athletes in their youth. These parents typically have higher expectations on their children and may unintentionally place pressure on them.

2 Data and Methodology

2.1 Data Collection and Transformation

To study swimmers' dropout behavior, it is essential to have accurate performance information, namely, swimming event times for each individual swimmer. SwimCloud.com, a well-known website among the swimming community, provides exactly such information. Table 2.1.1 below illustrates what the data looks like. We restrict our analysis only to adolescent swimmers in North Carolina from 2012 to 2024. The data covers 11,710 swimmers in 125 teams in 64 cities in North Carolina. Table 2.1.2 shows the summary statistics of our sample.

Table 2.1.1 Example of SwimCloud Data Structure

swimmer ID	Gender	Age Group	Swimmer Age	Meet Course	Meet ID	Meet Name	Meet Date	Event Name	Event Time
512252027	Men	13_14	13	SCY	195619	NC MAC Tar Heels States Meet	03/21/2021	100 Y Fly	1:27.09
512252027	Men	13_14	13	SCY	195619	NC MAC Tar Heels States Meet	03/21/2021	200 Y Free	2:31.65
512252027	Men	13_14	13	SCY	195619	NC MAC Tar Heels States Meet	03/21/2021	100 Y Back	1:20.73
512252027	Men	13_14	13	SCY	195619	NC MAC Tar Heels States Meet	03/21/2021	200 Y Back	2:49.89
512252027	Men	13_14	13	SCY	195619	NC MAC Tar Heels States Meet	03/21/2021	100 Y Free	1:09.72
512252027	Men	13_14	13	SCY	195619	NC MAC Tar Heels States Meet	03/21/2021	200 Y IM	2:53.63
512252027	Men	13_14	13	SCY	195619	NC MAC Tar Heels States Meet	03/21/2021	100 Y Breast	1:23.44
512252027	Men	13_14	13	SCY	195619	NC MAC Tar Heels States Meet	03/21/2021	200 Y Breast	3:05.49

Table 2.1.2 Summary Statistics

Number of Swimmers	Men	4,915
	Women	6,795
Swimmer Age (year)	min	4
	avg	13.15
	max	18
Number of Teams		125
Number of Meets		12,100
Number of City		64
Modeling Period	Start	5/1/2012
	End	12/31/2023

Swimmers typically swim multiple events in a meet. Different swimmers are good at different events. Thus, event times across different events are not directly comparable, even for the same swimmer. To address this, an index is created to standardize individual event times, allowing for fair comparisons of swimmers' performances.

A typical standardized scoring system compares each swimmer's event times to some benchmark times for that event. We choose to use the top 5 percentile of each event as the benchmark for that event. As an example, Table 2.1.3 shows the benchmark times for short course events.

Table 2.1.3 5th Percentile by Age Group (Short Course)

Event	Men						Women					
	Under 8	9 and 10	11 and 12	13 and 14	Over 15	Time Standard	Under 8	9 and 10	11 and 12	13 and 14	Over 15	Time Standard
50 Y Back	41.30	35.04	30.59			30.59	42.59	35.68	31.37			31.37
50 Y Breast	47.22	40.40	34.30			34.30	48.57	40.80	35.45			35.45
50 Y Fly	38.93	32.91	28.87			28.87	40.43	33.77	29.53			29.53
50 Y Free	35.57	30.05	26.46	23.52	21.77	21.77	36.92	30.82	27.32	25.53	24.47	24.47
100 Y Back	85.89	74.54	64.66	56.97	52.04	52.04	88.22	75.86	66.39	61.15	57.94	57.94
100 Y Breast	98.70	86.25	73.44	64.42	58.91	58.91	102.23	87.28	76.12	70.20	66.37	66.37
100 Y Fly	83.77	71.42	62.38	55.54	51.43	51.43	87.65	72.74	64.51	60.19	57.30	57.30
100 Y Free	78.59	65.84	57.36	51.18	47.47	47.47	81.76	67.37	59.09	55.26	52.98	52.98
200 Y Free	161.58	140.52	123.32	110.69	102.74	102.74	165.56	143.65	127.32	118.88	113.43	113.43
200 Y Back			131.58	121.21	111.83	111.83			136.21	130.17	123.95	123.95
200 Y Breast			149.80	138.14	126.99	126.99			154.80	148.86	141.67	141.67
200 Y Fly			132.28	120.22	112.02	112.02			136.00	129.87	123.65	123.65
200 Y IM			137.80	124.82	114.96	114.96			142.21	134.47	127.99	127.99
400 Y IM			280.69	258.82	240.85	240.85			291.06	276.08	265.26	265.26
500 Y Free			325.41	295.42	277.14	277.14			335.41	314.23	301.83	301.83

We further construct a universal benchmark time for each event, which is the fastest of all the benchmark times across all age groups, shown in column "Time Standard" in Table 2.1.3. The performance index at any point in time is defined as:

$$Performance\ Index_{it} = \frac{\sum_j^n Best\ Time_j}{\sum_j^n Event\ Time_{ijt}}$$

Where $Performance\ Index_{it}$ is the performance index for swimmer i in month t

$Best\ Time_j$ is the best time for event j to which swimmer i participated in month t

$Event\ Time_{ijt}$ is the individual time of swimmer i for event j in a month t

n is the total number of events participated by the swimmer i in month t

Performance index reflects a relative position of a swimmer to the top 5th percentile, a direct measurement of their performance. The lower the performance index value, the faster the swimmer is.

We define the end of a swimmer's career as the month when the swimmer swims their last meet. Swimmers can exit our dataset for several reasons, including turning 18 and going off to college. This kind of exit cannot be considered as quitting the sport. Therefore, a swimmer's dropout is defined as the month of their last meet, given that the swimmer is no older than 18.

$$\text{Dropout Indicator} = \begin{cases} 1 & the\ swimmer's\ last\ meet\ and\ swimmer's\ age \leq 18 \\ 0 & other\ meets \end{cases}$$

2.2 Methodology: Survival Function and Hazard Function

Survival analysis is a statistical method to predict the probability of an event occurring at certain times. In our case, the event is a swimmer's dropout (or quit) from swimming. Since one of our goals is to understand the probability of a swimmer quitting swimming, survival analysis provides an ideal framework.

A survival function defines the probability of a swimmer surviving up to age t. The probabilities of surviving from one age to another may be multiplied together to give the cumulative survival probability. Formally, it can be written as

$$S(t) = \prod_{t_i < t}\left(1 - \frac{d_{ti}}{n_{ti}}\right)$$

Where $S(t)$ is the survival probability at age t

n_t the number of swimmers who do not quit before t

d_t the number of swimmers who quit at t

The Cox proportional hazards model [8] is the most commonly used multivariate approach to estimate survival probability. As a regression model, it enables us to incorporate a set of covariates to forecast swimmers' dropouts as expressed by the hazard function. Mathematically, the Cox model is written as

$$h(t) = h_0(t)\, e^{\{\beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \ldots + \beta_p x_p\}}$$

where the hazard function $h(t)$ is determined by a set of covariates $(x_1, x_2, \ldots, x_p)$, whose impact is measured by the size of the respective coefficients $(\beta_1, \beta_2, \ldots, \beta_p)$. The term h_0 is called the baseline hazard, the value of the hazard if all the x_i are equal to zero.

3 Model Results

With a survival analysis framework outlined in Section 2, we aim to understand how swimmers' performance and their changes in relative positions in the state influence their decisions to quit. Figure 3.1.1 shows the Kaplan-Meier survival curve [9]. It naturally provides us with empirical survival probability (as well as dropout probability) by age. As we can see, swimmers' dropouts are related to swimmers' ages. Most dropouts occur around 13-14 age. Another observation is that we find men and women have different dropout rates. Girls' dropout rates are consistently higher than boys' rates starting from age 13.

3.1 Effects of Swimmer Self-Improvement

How our performance index is constructed makes it an effective way to measure a swimmer's performance progression. One way to capture a swimmer's performance index progression is to run a simple linear regression of the index on how long the swimmer has swum, measured in months. The slope of this regression tells us how fast the swimmer makes progress over time. The y-intercept of this regression shows the starting performance level of the swimmer, and the x-intercept shows the latest performance level of the swimmer. Intuitively, the latest performance should matter the most to swimmers' dropout decisions. We select the performance of the last three years, either prior to quitting or before the sample ends, for regression analysis. The linear regression is specified as:

$$Performance\ Index_i = \alpha_i + \beta_i number\ of\ swimming\ month_i$$

Where $Performance\ Index_i$ is the last three year performance index for swimmer i
$number\ of\ swimming\ month_i$ is number of swimming month for swimmer i

Figure 3.1.1 Kaplan-Meier Survival Curve by Gender

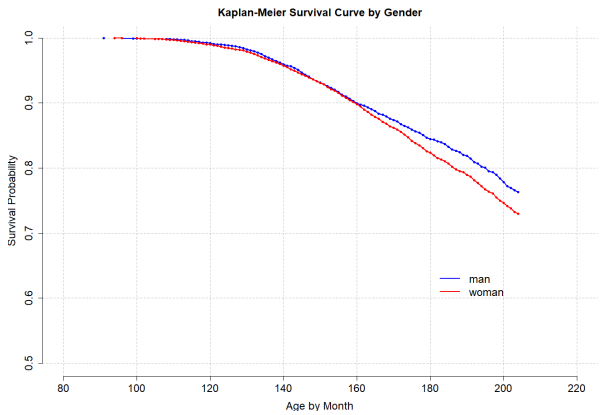
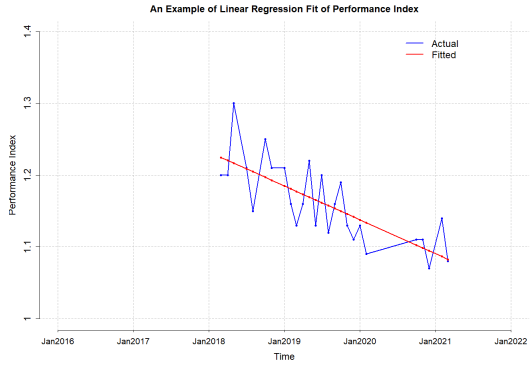


Figure 3.1.2, using a randomly selected swimmer as an example, shows how a simple linear regression fits to the performance index. Since swimmers typically become faster as they get older, the slope is typically negative.

Figure 3.1.2 Linear Regression Fit of Performance Index



Given we have obtained the slope and the intercept for each swimmer, the Cox hazard model can be specified as:

$$h(t) = h_0(\mathrm{t})\, e^{\{\beta_1 Woman\_Ind + \beta_2 Intercept + \beta_3 Slope\}}$$

Where Woman Ind $= \begin{cases} 1 & the\ swimmer\ is\ a\ woman \\ 0 & otherwise \end{cases}$

Intercept is the intercept from performance index regression for each swimmer
Slope is the slope from performance index regression for each swimmer

The estimation results are shown in Table 3.1.1.

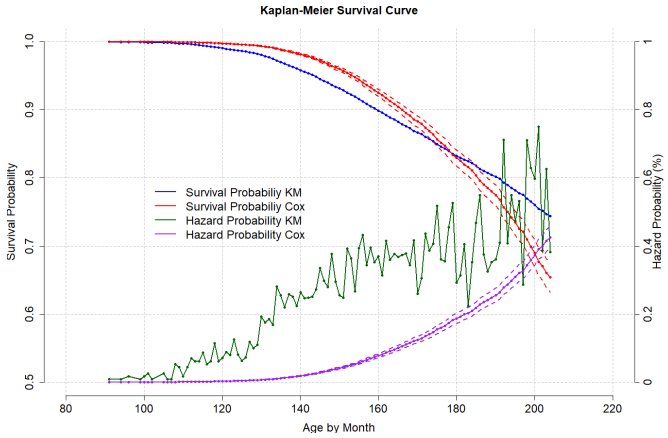
Table 3.1.1 Cox Hazard Model Estimation Results

Variable	coef	exp(coef)	se(coef)	z	Pr(>|z|)	lower 0.95	upper 0.95
Women Ind	0.58	1.78	0.05	12.38	<2e-16	1.63	1.95
Intercept	4.46	86.39	0.09	49.42	<2e-16	72.39	103.10
Slope	3.29	26.82	0.35	9.51	<2e-16	13.62	52.83

There are several observations from the table above: First, the coefficient of the indicator for "woman" is positive and statistically significant, suggesting that if you hold other variables constant, female swimmers are more likely to quit than male ones. Second, the coefficient of the intercept is also positive. The intercept variable reflects a swimmer's general performance level. The larger the intercept, the larger the swimmer's general performance index value. As discussed earlier, a larger performance index value implies that the swimmer is slower relative to other swimmers. Thus, the positive coefficient of the intercept implies that a swimmer's absolute speed matters. The slower a swimmer's absolute speed is, the more likely the swimmer will quit. Third, the coefficient of slope is positive as well. The slope indicates how fast a swimmer is making progress over time. When a swimmer is making good progress, the downward slope will be steeper and its value will be more negative. If a swimmer does not make good progress, the slope will be less negative, close to 0, or even positive. Since the coefficient of slope is positive, a less negative value will generate higher hazard rate than a more negative slope value. Thus, when a swimmer is making slower or no progress, the swimmer is more likely to quit.

Figure 3.1.3 shows the survival curve and hazard curve (with 90% confidence interval) predicted by the Cox hazard model and the Kaplan-Meier curve together.

Figure 3.1.3 Survival Curve by Cox Hazard Model



3.2 Effects of Swimmers' Relative Position

In this section, we study a swimmer's relative position compared with peers. How do we categorize a swimmer as fast or slow? Within each age group, we divide swimmers into two groups based on how their times in any event compared with the 50th percentile of that event. If a swimmer with any event in the top (or bottom) 50th percentile in his or her age group, the swimmer will be categorized as fast (or slow) swimmer. A swimmer can be a fast swimmer in some age group(s), but a slow swimmer in other age groups.

We are interested in the case where a swimmer was faster when they were young, but becomes slower when they grow up. For example, a swimmer was in the faster group when he was younger than 13. Over the years, although they still make progress, they can no longer make it to the fast group when they are older than 13 years old. We are interested in how this kind of relative position change influences a swimmer's decision to quit. To articulate this kind of relative position change, we construct a series of variables. First, we define the following indicator variables.

$$\text{Fast 15 ind} = \begin{cases} 1 & \textit{the swimmer is in top } 50\ \%\ \textit{ percentile in age group } 15-18 \\ 0 & \textit{the swimmer is not in top } 50\ \%\ \textit{ percentile in age group } 15-18 \end{cases}$$

$$\text{Fast 13 ind} = \begin{cases} 1 & \textit{the swimmer is in top } 50\ \%\ \textit{ percentile in age group } 13-14 \\ 0 & \textit{the swimmer is not in top } 50\ \%\ \textit{ percentile in age group } 13-14 \end{cases}$$

$$\text{Fast 11 ind} = \begin{cases} 1 & \textit{the swimmer is in top } 50\ \%\ \textit{ percentile in age group } 11-12 \\ 0 & \textit{the swimmer is not in top } 50\ \%\ \textit{ percentile in age group } 11-12 \end{cases}$$

$$\text{Fast 09 ind} = \begin{cases} 1 & \textit{the swimmer is in top } 50\ \%\ \textit{ percentile in age group } 9-10\ \textit{ or younger} \\ 0 & \textit{the swimmer is not in top } 50\ \%\ \textit{ percentile in age group } 9-10\ \textit{ or younger} \end{cases}$$

To capture each swimmer's relative position changes, we define another set of indicators.

$$\text{Top50 15 ind} = \begin{cases} 1 & \text{Fast 15 ind} = 1 \\ 0 & \text{Fast 15 ind} = 0 \end{cases}$$

$$\text{Top50 13 ind} = \begin{cases} 1 & \text{Fast 13 ind} = 1 \text{ and Fast 15 ind} = 0 \\ 0 & \text{otherwise} \end{cases}$$

$$\text{Top50 11 ind} = \begin{cases} 1 & \text{Fast 11 ind} = 1 \text{ and Fast 13 ind} = 0 \text{ and Fast 15 ind} = 0 \\ 0 & \text{otherwise} \end{cases}$$

$$\text{Top50 09 ind} = \begin{cases} 1 & \text{Fast 09 ind} = 1 \text{ and Fast 11 ind} = 0 \text{ and Fast 13 ind} = 0 \text{ and Fast 15 ind} = 0 \\ 0 & \text{otherwise} \end{cases}$$

Using the changes in the relative position for each swimmer, the Cox hazard model can be specified as:

$$h(t) = h_0(t)\, e^{\{\beta_1 Women\_Ind + \beta_2 top50\_15\_ind + \beta_3 top50\_13\_ind + \beta_4 top50\_11\_ind + \beta_5 top50\_09\_ind\}}$$

Where Women Ind = $\begin{cases} 1 & \textit{the swimmer is a women} \\ 0 & \textit{otherwise} \end{cases}$

The estimation results are shown in Table 3.2.1.

Table 3.2.1 Cox Hazard Model Estimation Results

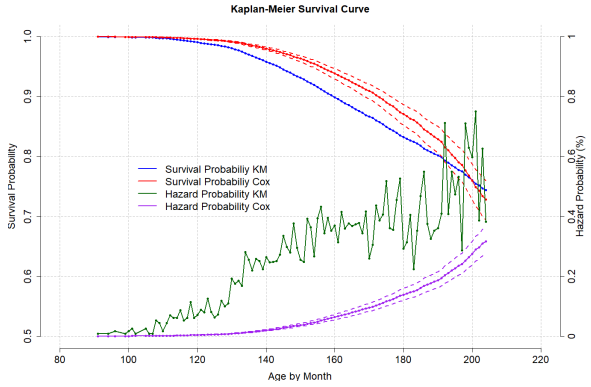
Variable	coef	exp(coef)	se(coef)	z	Pr(>|z|)	lower 0.95	upper 0.95
Women Ind	-0.04	0.96	0.04	-0.95	0.343	0.88	1.05
top50_15_ind	-2.75	0.06	0.09	-30.97	<2e-16	0.05	0.08
top50_13_ind	-0.55	0.58	0.06	-8.43	<2e-16	0.51	0.66
top50_11_ind	0.68	1.98	0.06	10.91	<2e-16	1.75	2.24
top50_09_ind	1.55	4.72	0.06	23.88	<2e-16	4.15	5.36

The coefficient of Top50_15_ind is negative and statistically significant, meaning that for swimmers older than 15 years old, being in the fast group makes them less likely to quit. Although the coefficient of Top50_13_ind is also negative, its magnitude is much smaller, but still statistically significant. That suggests that if swimmers are fast when they are in the 13 to 14 age group but become slow when in the 15 and older group, being in fast group earlier still makes them less likely to quit, but the effect is much smaller.

The coefficients of Top50_11_ind and Top50_09_ind are positive. That means swimmers who are only faster when they are younger are more likely to quit, especially for those who are faster when they are 9 years old.

Figure 3.2.1 compares the survival curve predicted by the Cox hazard model to Kaplan-Meier curve. An important observation is that, with more covariates, the Cox curve is closer to the Kaplan-Meier curve.

Figure 3.2.1 Survival Curve by Cox Hazard Model



4 Conclusion

This research aims to understand how swimmers' improvements influence their decisions to quit. By using the survival analysis method to analyze roughly 12,000 swimmers' meet records over 10 years, this research has several important findings. First, this analysis presents a precise estimation of swimmers' dropout probabilities at different ages. Although age has been identified as a factor influencing dropout in literature, this research is the first that carries out a precise estimation of dropout probabilities at different ages. Second, the research finds that a swimmer's absolute performance level and the speed of improvement both influence his or her dropout probability. The faster a swimmer is, the less likely the swimmer will quit. The faster the swimmer's improvement is, the less likely the swimmer will quit. But the absolute performance level is a more important factor. This finding makes sense. For many swimmers, one of the main goals of swimming is to meet cuts for selective meets. When a swimmer realizes that their times cannot reach those standards, the swimmer may consider quitting. Third, the research finds that swimmers who are faster when they are younger but slower when they grow up are

more likely to quit. Most swimmers become faster as they grow up, but if they are slow among peers, regardless how much progress they have made, they are more likely to quit. Finally, the research confirms a consensus observation in the swimming community that women are more likely to quit than men.

The research makes several contributions to the existing literature. First, the research establishes the importance of utilizing large and long historical data for analyzing sports related behavior. Second, the research deepens the understanding of adolescent dropout behavior by highlighting the importance of performance improvements, including both self-improvements and improvements on relative position among peers. Third, survival analysis is the main statistical framework for this research. Survival analysis is designed to understand the survival probability over treatment time. In this research, we focus on swimmers' natural age instead of years of swimming since we find using natural age has more meaningful implication. Thus, the research expands the usability of the framework.

References

[1] McClone, Nicole S., "Psychological Factors That Impact the Drop-out Rate in Adolescent Sports", HIM 1990-2015, 2015

[2] Monteiro, Diogo; Luis Cid; Daniel Almeida Marinho; Joao Moutao; Anabela Vitorino; Teresa Bento, "Determinants and Reasons for Dropout in Swimming – Systematic Review", Sports, Vol. 5, Iss. 3, 2017

[3] Fraser-Thomas, Jessica, "Examining Adolescent Sport Dropout and Prolonged Engagement from a Developmental Perspective", Journal of Applied Sport Psychology, 2008

[4] Raedeke, Thomas D. and Alan L. Smith, "Coping Resources and Athlete Burnout: An Examination of Stress Mediated and Moderation Hypotheses", Journal of Sport and Exercise Psychology, 2004.

[5] Salguero, A; R. Gonzalez-boto; C. Tuero; S. Marquez, "Identification of Dropout Reasons in Yound Competitive Swimmers", Journal of Sports Medicine and Physical Fitness, Vol. 43, Iss. 4, 2003

[6] Larson, Heather K.; Bradley W. Young; Tara-Leigh F Mchugh, Wendy. M. Rodgers, "A Multi-Theoretical Investigation of the Relative Importance of Training Volume and Coach Autonomy Support for Preventing Youth Swimming Attrition", Sport Science, 2020

[7] Larson, Heather K.; Bradley W. Young; I. Reade, "Exploring High School Swimmers' Sources of Sport Commitment", International Journal of Sport Psychology, Vol. 49, Iss. 2, 2018

[8] Cox, D. R., "Regression Models and Life-Tables", Journal of the Royal Statistical Society, Vol. 34, No. 2, 1972

[9] Kaplan, E. L. and Meier, P., "Nonparametric Estimation from Incomplete Observations", Journal of the American Statistical Association, Vol. 53, No. 282, 1958.