STATISTICA, anno LXXXIII, n. 4, 2023

PERFORMANCE OF INDIAN BATSMEN IN THE 2023 WORLD CUP SQUAD - A SURVIVAL ANALYSIS APPROACH

M. Ramakrishnan [1]
Department of Mathematics, Ramakrishna Mission Vivekananda College, Chennai, India
N. Viswanathan
Department of Statistics, Presidency College, Chennai, India
R. Ravanan
Joint Director (Planning and Development), Directorate of Collegiate Education, Chennai, India

SUMMARY

The game of cricket is enjoyed by millions of fans across the Globe. India is perusing the game like anything. Though, at present, many local tournaments like IPL were conducted in India, the ability of the batsmen and their patterns of scoring runs were reflected by the amount of runs they score in the international ODI's played between Countries. The runs scored by batsmen with information on whether they are "batting-first" or "chasing" throws more light on the patterns of runs-scoring. Also, the order in which the batsmen bat is also taken into consideration in this study. Further, the performance of batsmen varies across different teams of the game. More uncertainty exists in the run-scoring pattern. All these make it difficult for predicting the runs scored by a batsman. Survival analysis comes handy in predicting the probabilities of such events. The study, in this perspective, considers nine batsmen selected from the one-day Indian world cup squad 2023. The information of these batsmen, particularly the runs scored by them against different countries were used to find the probabilities of their run-scoring pattern. The study uses Kaplan-Meier's product limit estimator, Cox Proportional Hazard model and Accelerated Failure Time Parametric model for analysing the patterns. Log-rank test is used for comparing survival distributions. Also, the study compares the relative performance of selected batsmen. Data, updated as on 4th September 2023, were used for each of the batsman under consideration. The analysis has been carried out using R program.

Keywords: Survival analysis; Kaplan-Meier's estimate; Cox Proportional Hazard; Parametric Model.

[1] Corresponding Author. E-mail: mramkey@rkmvc.ac.in

1. Introduction

Cricket is a very popular sport throughout the world. This game is conducted across countries, states and even between cities now-a-days. The International Cricket Council (ICC) is the international governing body of cricket. 12 countries are full members of the ICC that also includes India. One-day cricket World cup is conducted once in every 4 years. The first ODI was played on 5^{th} January 1971 between Australia and England at Melbourne Cricket Ground. 13^{th} ICC Men's World Cup 2023 conducted by ICC from 5^{th} October to 19^{th} November 2023 in various stadiums across India. The teams from Australia, England, Pakistan, India, Bangladesh, Sri Lanka, Afghanistan, New Zealand, South Africa and Netherlands participated in this tournament. India won ICC ODI World Cups in 1983 and 2011, beating West Indies and Sri Lanka in the finals, under the captaincy of Kapildev and Dhoni, respectively.

Winning Chance of every team depends on the Performance of batsmen, bowlers, and fielders in each team. Of the three, in this study, the focus is on batting performance. In the recent past, many research works have been carried out in the direction of predicting batmen's performance based on their past records.

Survival Analysis is the study of time-to-event. It differs from other branches of Statistics mainly because of the incomplete information that arise due to censoring. It considers the response variable as the duration of time to a specific pre-defined event. In the case of batsmen, the runs scored by him is taken as the length of his survival and the terminal event is taken as his dismissal from the crease. The performance, in terms of the runs becomes incomplete or censored, if the match ends with or without a success for his team.

The other covariates that need to be controlled are the status of 'batting-first' or 'chasing', the order in which a batsman is allowed to enter batting, the strength of the opposition team and the nature of the pitch being played. Of the four, this study focusses only on the first two factors as controls. Stevenson and Brewer (2017) modelled the hazard function for batsmen in Test Cricket using Bayesian approach and identified that the batsmen's speed of transition from initial state to an equilibrium state serves as an indicator of their performance measured in terms of number of runs. Shah and Patel (2018) ranked the captains of the teams in ODI based on several parameters using Principal Component Analysis and, they showed how to rank captains based on their individual contributions to their teams and the team's performance under their captaincy. Kalpdrum and Nirav Kumar (2018) suggested that Random Forest method turned out to be the most accurate classifier for predicting how many runs a batsman is likely to score and how many wickets a bowler is likely to take in a ODI match and this will help the team management select best players for the team. Kachoyan and West (2018) have derived a recurrence relation for batsmen's survival function using the probability of being dismissed at each score when considering their test matches played between Australia and India from 2008 to 2016. Manoj Ishi and Patil (2022) identified Logistic Regression and Support Vector Machine give better results compared to other models for predicting the winner of ODI Matches. Many researchers have focused their study on performance

of batting and their predictions. Mohan *et al.* (2022) applied Weibull smoothening and fuzzy linear regression approach to estimate survival probabilities that varies over an interval, compared to the Kaplan-Meier's constant probabilities of survival in the intervals. Shah *et al.* (2023) studied about the survival probabilities of top 10 ODI batsmen around the world and it can be used as a new measure for evaluating batsmen as it gives the ability of the batsman to survive on crease. Preetham *et al.* (2023) suggested a model for predicting the results of the IPL matches, in particular, forecast the score of an innings using machine learning models.

In this paper, we have taken past records of nine batsmen selected from Indian squad for ICC Men's World Cup 2023. This data in conjunction with Kaplan-Meier nonparametric model, Cox PH semiparametric model and parametric model were used to estimate different measures of performance of batsmen, the factors influencing it and the distribution that best suits for a batsman's run scoring pattern. Using Kaplan-Meier model, the probabilities of batsman scoring specific runs are estimated and these probabilities are compared across different batting order and that of batting-first or chasing. Individual batsman's performance is compared between "batting-first" and "chasing". After grouping the batsmen into two groups, one designated as top-order and the other as middle-order, log-rank test is used to compare these two groups. A separate comparison among the top-order and that of the middle-order batsmen is also carried out using log-rank test.

Further, conditional probabilities of a batsman scoring a specific number of runs, given that he has already scored a specific number of runs were also estimated. To study the impact of status of batting, namely, "batting-first" or "chasing" and that of the batting order of a batsman, these are included as covariates in the Cox PH model and analyzed using their respective hazard ratios. Parametric models are used to identify the best-fit distribution for the probabilities of scoring runs by each batsman. The criteria of AIC and BIC were used to identify the distribution of best-fit. The factors influencing these probabilities are further verified using time ratios under Accelerated Failure Time (AFT) parametric model. These metrics would help us to find the chance of a batsman scoring well in a given game, which in turn will help us constitute a better team that maximizes the chance of India's win in the cricket arena.

2. Methodology

Survival Analysis is the study about time-to-event data and Survival models mostly characterizes the probability of survival using the incomplete data, also called the censored data. This Paper uses Kaplan-Meier's method of estimation, Cox PH model and Parametric Model for deriving and comparing results concerning Batsmen's performance.

2.1. *Kaplan-Meier's Model*

Survival probabilities were estimated through the product of conditional probabilities without assuming distributional form for survival time (Kaplan and Meier, 1958). In this model, Survival Function $S(t) = P(T > t)$ is estimated through

$$\widehat{S}(t) = \prod_{t_i \leq t} (1 - \frac{d_i}{n_i}),$$

where d_i and n_i respectively are the number of events that occur and the number of subjects that enters the study at time t_i, which is the i^{th} ordered survival time. Log-rank test is used to compare the survival patterns between "batting-first" and "chasing" for individual players. All top-order batsmen taken as a single group is compared with another group consisting of middle-order batsmen, using log-rank test. Within these two groups, individual comparisons were also made using log-rank test. Kaplan-Meier Survival probabilities of selected batsmen were compared against chosen countries. Conditional Survival probabilities for each batsman are calculated through Kaplan-Meier's method.

2.2. *Cox Proportional Hazard Model*

Cox (1972) proposed the following regression model for the hazard function

$$h(t|X, \beta) = h_0(t) e^{\sum_{i=1}^{p} \beta_i X_i},$$

where:

- t represents the survival time;
- $h(t|X, \beta)$is the hazard function determined by a set of p covariates $(X_1, X_2, ..., X_P)$;
- The coefficients $(\beta_1, \beta_2, ..., \beta_p)$ measure the impact of covariates;
- The term $h_0(t)$ is called the baseline hazard.

2.3. *Proportional Hazard Assumption*

The proportional hazards assumption requires that covariates are multiplicatively related to the hazard. To verify Proportional hazard assumption, test based on Schoenfeld Residuals is used.

Cox PH model, in this paper, is used with two categorical covariates. The first covariate "Innings" is dichotomized with "batting-first" and "chasing". The second covariate "Batting order" has three categories, namely, "top-order", "middle-order" and "low-order". Our Cox model, in this context, does not use the interaction effect. A separate Cox PH model was developed for comparing the hazard ratios all batsmen against Kohli, taken as base line reference.

2.4. Parametric Models

If the survival time is assumed to follow some specific distribution, then Parametric models can be applied. In this paper, Accelerated Failure Time approach is used to obtain the acceleration factor and comparison is made in the scoring patterns between the "batting-first" and "chasing". AIC and BIC measures are used to identify the distribution of the best-fit among Exponential, Weibull, Log-Normal and Log-Logistic distributions.

3. DATA STRUCTURE

The data on the performance of World Cup 2023 Indian team Batsmen was taken from www.espncricinfo.com. The data considered in this study includes all ODI matches played up to September 4, 2023 by the Batsmen Kohli, Rohith, Rahul, Gill, Shreyas Iyer, Pandia, Ishan, Jadeja, and Suryakumar. Event of interest of this study is 'getting out' in an innings and considering "not out" as censored ones. Kohli, Rohith and Jadeja played 266, 239 and 122 matches respectively but the others played less than 60 matches, with Ishan playing the minimum, 17 matches.

In this paper, runs scored by the Batsmen were considered as their survival time. Nonparametric Kaplan-Meier, Semiparametric Cox-PH and Parametric Models were used to study the survival pattern present with respect to "innings" (Batting first or Chasing) and Positions (Top, Middle and Low). Kaplan-Meier estimates are compared using Log-rank test. Comparing all the batsmen simultaneously may not be justified due to high degree of heterogeneity present in the batting condition and style and as such batsmen within "top-order" and "middle-order" were compared among themselves. A two-group comparison was also carried out to bring out the differences present between top and middle-order batsmen. The low-order batsmen were not considered, as they are highly likely to be bowlers.

In the context of this paper, batsmen, playing predominantly in positions one, two and three are taken as "top-order" batsmen and those playing in positions four, five, six and seven are considered as "middle-order" batsmen and the remaining positions are considered as "low-order".

Applying this arrangement, we have Kohli, Rohit, Rahul, Gill and Ishan considered as "top-order" batsmen and Shreyas Iyer, Pandia, Jadeja and Suryakumar as the "middle-order" batsmen.

Survival probabilities of selected batsmen against six chosen countries are studied. For this comparison, those batsmen who have played a minimum of fifty matches in total were considered. If a particular batsman has played at least ten matches against a country, that particular combination of batsman-country is considered for estimation of survival probabilities. These estimates were derived using Kaplan-Meier Model.

In the game of cricket, at different stages of batting, we need an approximate measure for a batsman to achieve a score, given that he has already at a particular score. This measure will help us to decide the order in which the batsmen could be brought into the batting sequence. Conditional probability is the required measure and this is estimated

using Kaplan-Meier Model. For this model, scores at 10,30,50,70 and 100 are considered.

4. Empirical Analysis

The details of the runs scored by the selected batsmen are presented for overall comparison in Table 1.

From Table 1, it is seen that, Kohli has played the maximum number of matches with 266 and Gill has the highest average of 52.2 runs. The lowest average is 21.09, that is concerned with Jadeja.

TABLE 1
Descriptive statistics of the selected batsmen (= Not out).*

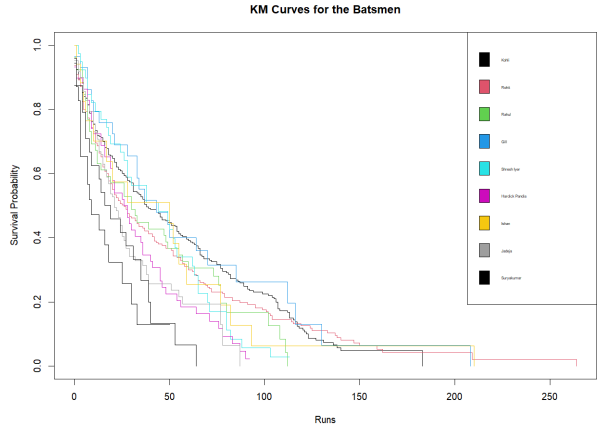
Player	Number of Matches	Min. Runs	Q1	Median	Mean	Q3	Max. Runs	SD
Kohli	266	0	10	35.50	48.50	79.75	183	43.13
Rohit	239	0	8	22.00	41.51	63.00	264	45.90
Rahul	52	0	8	28.00	38.19	63.25	112	34.67
Gill	29	0	20	40.00	52.21	70.00	208	46.80
Ishan	17	1	8	28.00	45.65	59.00	210	52.08
Shreyas Iyer	39	2	17.5	38.00	42.18	64.00	113*	29.73
Pandia	59	0	9	21.00	29.71	43.00	92*	25.82
Jadeja	122	0	7	16.00	21.10	26.75	87	19.11
Suryakumar	24	0	6	17.50	21.29	34.00	64	17.62

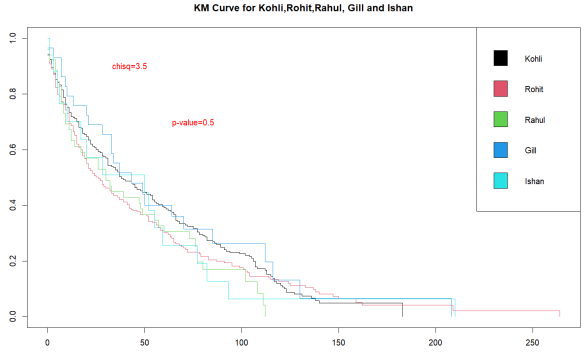
4.1. *Nonparametric Model*

Survival in this case means 'not getting out' while batting. In this study, hazard represents 'getting out'. Kaplan-Meier (KM) analysis is nonparametric in nature and in this study, only the event, "the batsmen getting out" and the time, "Runs scored by the batsmen" are used in the analysis. Survival probabilities of the batsmen using KM estimator is presented in Table 2 and its graph is shown in Figure 1.

TABLE 2
KM survival probabilities of the batsmen (n = number of matches played).

Runs (Time)	Kohli (n=266)	Rohit (n=239)	Rahul (n=52)	Gill (n=29)	Shreyas Iyer (n=39)	Pandia (n=59)	Ishan (n=17)	Jadeja (n=122)	Suryakumar (n=24)
10	0.7538	0.7296	0.6923	0.7931	0.8205	0.7233	0.7010	0.7537	0.6250
30	0.5705	0.4617	0.4891	0.6552	0.5629	0.4241	0.5098	0.3343	0.3750
50	0.4481	0.3669	0.3668	0.4002	0.4258	0.2253	0.4461	0.2480	0.1333
70	0.3339	0.2468	0.3057	0.3152	0.1987	0.1639	0.2549	0.1873	
100	0.2256	0.1754	0.1681	0.2627	0.0568		0.0637		





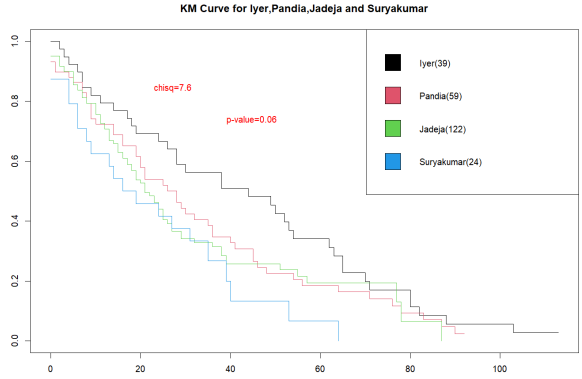


Figure 1 – Survival curves comparing all batsmen, top-order and middle-order Batsmen.

From Table 1 , it is seen that Kohli, Rohit, Rahul and Gill have high probability of getting good runs. Also, it reveals that Kohli has nearly 23% chance for hitting a century and 45% chance for hitting a half-century. Similarly, we expect Gill, Shreyas Iyer, and Ishan to have a good chance of scoring 50 or more runs ($S(50) > 0.4$). Also, the survival probabilities reveal that Pandia, Jadeja, and Suryakumar used to get a good start but are unable to convert them into a big score. The last four batsmen have considerably low probability of scoring a century and it may be due to their low-order appearance in the batting order. On the contrary, Kohli and Gill are expected to score more centuries ($S(100) > 0.2$). From Figure 1, it is observed that the batsmen Rohit, Shreyas Iyer, and Ishan have scored more than 200 runs.

On considering Kohli, Rohit, Rahul, Gill and Ishan among the top-order batsmen (Figure 1), it is observed that there exists no significant difference among these batsmen in their run scoring pattern ($\chi^2 = 3.5$, p-value > 0.05). But, while considering Shreyas Iyer, Pandia, Jadeja and Suryakumar in a single group as middle-order batsmen (Figure 1), it is found that there is a significant difference among their run scoring patterns at 10% level of significance ($\chi^2 = 7.6$, p-value $= 0.06$).

Run scoring pattern of batsmen within the top-order batsmen is compared using log-rank test and the results are presented in Table 3. Similarly, the run scoring pattern of batsmen within the middle-order group is compared and the results are presented in Table 4.

TABLE 3
Comparison of run scoring pattern among top-order batsmen (a/b: "a" denotes chi-square value and "b" denotes the associated p-value).

Player	Kohli	Rohit	Rahul	Gill	Ishan
Kohli		1.666/0.197	2.683/0.101	0.347/0.556	0.248/0.618
Rohit			0.174/0.676	1.012/0.314	0.002/0.964
Rahul				2.800/0.095	0.018/0.894
Gill					0.430/0.512

On comparing the batsmen within the top-order, we find no significant difference in the run scoring pattern (p-values> 0.05 in Table 3). On comparing the batsmen within the middle-order group, we find significant difference among the batsmen, in particular, between Shreyas Iyer and that of Jadeja and Suryakumar at 5% level (Table 4).

When comparing the run scoring pattern between the top and middle-order groups, it is found that there is a significant difference ($p < 0.001$) between the two groups and it is further observed that top-order batsmen score considerably high number of runs compared to the middle-order batsmen.

The difference between “batting-first” and “chasing” is studied using log-rank test for each of the batsman separately and the results are presented in the Table 5.

From Table 5, it is observed that, except for the player Kohli, for all other bats-

TABLE 4
Comparison of run scoring patterns among middle-order batsmen (a/b: "a" denotes chi-square value and "b" denotes the associated p-value).

Player	Shreyas Iyer	Pandia	Jadeja	Surya Kumar
Shreyas Iyer		2.031/0.155	4.346/0.037	8.501/0.004
Pandia			0.237/0.626	2.699/0.100
Jadeja				1.374/0.241

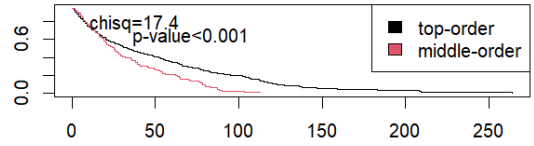


Figure 2 – KM Curves for top and middle-order batsmen.

TABLE 5
Difference between "batting-first" and "chasing" for individual batsman.

Player	Innings	10	30	50	70	100	p-value
Kohli	I(n= 120,45.98) II(146, 50.57)	0.6967 0.8003	0.5346 0.5998	0.4299 0.4627	0.2895 0.3718	0.1822 0.2650	0.0506
Rohit	I (103, 44.33) II(136,39.38)	0.6783 0.7689	0.4293 0.4870	0.3392 0.3887	0.2438 0.2490	0.1773 0.1744	0.5812
Rahul	I (26,44.15) II (26,32.23)	0.6923 0.6923	0.5701 0.4072	0.4072 0.3258	0.3258 0.2851	0.1862 0.1425	0.4418
Gill	I (17, 63.76) II (12,35.83)	0.7647 0.8333	0.7059 0.5833	0.5229 0.2083	0.3922 0.2083	0.3268	0.2573
Shreyas Iyer	I(24,40.25) II(15,45.33)	0.7917 0.8667	0.5000 0.6667	0.3750 0.5079	0.2083 0.1693	0.0417 0.0847	0.3899
Pandia	I(34,27.05) II(25,33.32)	0.6471 0.8343	0.3379 0.5562	0.2027 0.2649	0.1351 0.2119		0.3820
Ishan	I (7,70.85) II(10,28)	0.8571 0.5833	0.7143 0.3500	0.5714 0.3500	0.4286 0.1167	0.1429	0.2394
Jadeja	I(63,19.94) II(59,22)	0.7229 0.7851	0.3810 0.2954	0.2857 0.2167	0.2084 0.1733		0.6833
Suryakumar	I(14,20.21) II(10,22.8)	0.5714 0.7000	0.3571 0.4000	0.0952 0.2000			0.6598

men, no significant difference exists between the runs scored while "batting-first" and "chasing" (p-value > 0.23). For player Kohli, the associated p-value ($p = 0.0506$) is very close to 0.05 and hence it can be concluded that there exists significant difference in runs scored while "batting- first" and "chasing" at 6% level of significance. This difference is consistently seen from the run-scoring probability of Kohli at the selected cutoff values, 10, 30, 50, 70 and 100 runs. Certain useful patterns emerge from Table 5. Though, there are no statistical significance, it is observed that batsmen Rohith, Pandia and Suryakumar scored more runs while chasing, whereas batsmen Rahul and Ishan score more runs while "batting-first". Batsmen Jadeja and Gill have a mixed pattern, in that they have high probabilities for scoring more number of runs in "batting-first" compared to "chasing" and for lesser number of runs, the trend is reversed.

Batsmen with minimum 50 matches played and those countries against which they played at least 10 matches are considered for estimating probabilities of scoring 10, 30, 50, 70 and 100 runs. These results are presented in Table 6.

TABLE 6
Survival probabilities of five batsmen against six teams (NOM = Number of matches played).

Player		Countries	Pakistan	Australia	England	New Zealand	South Africa	Sri Lanka
Kohli	Runs	NOM	14	44	35	29	28	49
		10	0.5000	0.7925	0.7078	0.6897	0.8214	0.7731
		30	0.3571	0.5594	0.5309	0.5862	0.6429	0.6268
		50	0.3571	0.4429	0.3539	0.4483	0.4643	0.4550
		70	0.3571	0.3263	0.2065	0.3035	0.3095	0.3212
		100	0.2679	0.1865	0.1475	0.2276	0.2257	0.2718
Rohit	Runs	NOM	17	42	19	25	24	48
		10	0.8157	0.7612	0.6802	0.7200	0.6667	0.6374
		30	0.5020	0.5893	0.3710	0.4400	0.2500	0.4385
		50	0.5020	0.4125	0.3092	0.3150	0.2083	0.3167
		70	0.1882	0.3094	0.2473	0.1800	0.1250	0.2375
		100	0.1255	0.2063	0.1649	0.0900	0.1250	0.1781
Rahul	Runs	NOM	1	11	9	5	4	7
		10		0.8180				
		30		0.5110				
		50		0.3070				
		70		0.3070				
		100						
Jadeja	Runs	NOM	7	28	20	10	6	17
		10		0.7263	0.7841	0.8000		0.7466
		30		0.1187	0.4621	0.4571		0.5973
		50		0.1187	0.3466	0.4571		0.2986
		70			0.3466	0.3429		
		100						
Pandia	Runs	NOM	4	11	11	10	5	10
		10		0.909	0.909	0.8		0.467
		30		0.636	0.606	0.4		0.114
		50		0.364	0.379	0.1		
		70		0.364	0.126			
		100						

Against Pakistan, the probability of scoring 30 and 50 runs is high for Rohith. When

it comes to 70 and 100 runs, Kohli has the highest probability of scoring. Thus, we see that to play against Pakistan these two players are inevitable. Against Australia, the probability of scoring 30 or 70 runs is high for Pandia. When it comes to 50 runs, Kohli has the highest probability of scoring, and for 100 runs Rohith has the highest probability. Thus, In case of Australia, the players we are looking for are Pandia, Kohli and Rohith. In a similar way, we found Rohith and Pandia suitable for England, Kohli and Jadeja for New Zealand and Kohli for both South Africa and Sri Lanka. These results are also reflected in Figure 3, depicting the survival curves of the selected players against the chosen countries.

Kaplan-Meier survival probabilities are used to find the conditional probability of a player scoring a specified number of runs, having known that he has already crossed a particular score. These conditional probabilities are estimated and are listed in Table 7.

This conditional probability, presented in the Table 7, represents the chances for a particular batsman to score 'b' runs or more, when he is playing at 'a' runs. This predicts the ability of each batsman in converting their score to a higher one, during the on-going play. The players listed in Table 7 are highly heterogeneous with respect to the number of matches they have played. Senior players such as Rohit and Kohli have played 266 and 239 matches respectively, whereas, players Gill and Ishan have played 29 and 17 matches respectively. These variations in the number of matches played should be taken into account while comparing the conditional probabilities.

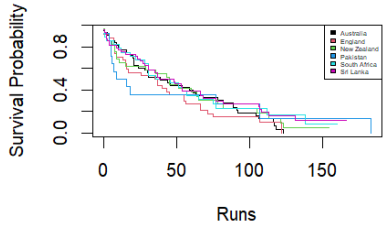
Playing at 30 runs, the chance of scoring 50 is more for Ishan, followed by Rohith and Kohli, in that order. When the number of matches played is taken into account, the performance of Rohith and Kohli stands tall compared to Ishan. Playing at 50 runs, the chance of scoring 100 is more for Gill, followed by Kohli and Rohith, in that order. Again, taking in to account, the number of matches played, the performance of Rohith and Kohli stands apart. Playing at 70 runs, the chance of scoring 100 is more for Gill, followed by Rohith and Kohli, in that order, to be interpreted in the same lines as in the previous two cases.

These conditional probabilities, in addition to the information about the number of matches played could help placing a particular player in a particular batting slot, for a given score of the team in a particular over. This will further help the team management to suggest the sequence in which the players should be sent into bat in a given scenario.

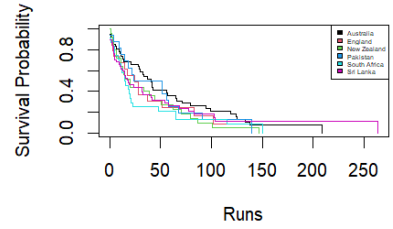
4.2. Semiparametric Model

In this section, a set of seven Cox PH models, one for each batsman were used to study the impact of "batting-first" or "chasing", represented by the variable "innings" and that of the position in which a batsman comes into bat.Another Cox PH model was used to compare the rest of the batsmen's performance with that of Kohli. In this case, estimated hazard ratios were used for comparison.

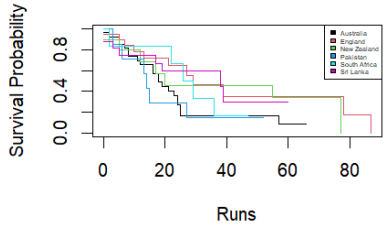
The results concerned with studying the impact of "innings" and "batting order" are presented in Table 8 and the verification of the PH assumption for Kohli is presented in



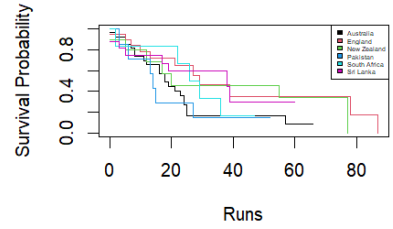
(a) Kohli



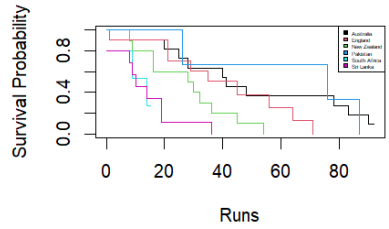
(b) Rohith



(c) Raul



(d) Jadeja



(e) Pandia

Figure 3 – Survival curves of batsmen against different countries.

TABLE 7
Conditional probabilities $P(Runs > b | Runs > a)$ ($^+$denotes number of matches played).

Player	b	a = 10	a = 30	a = 50	a = 70
Kohli 266$^+$	10	1.0000			
	30	0.7568	1.0000		
	50	0.5945	0.7855	1.0000	
	70	0.4430	0.5853	0.7451	1.0000
	100	0.2993	0.3954	0.5035	0.6757
Rohit 239$^+$	10	1.0000			
	30	0.6328	1.0000		
	50	0.5029	0.7947	1.0000	
	70	0.3383	0.5345	0.6727	1.0000
	100	0.2404	0.3799	0.4781	0.7107
Rahul 52$^+$	10	1.0000			
	30	0.7065	1.0000		
	50	0.5298	0.7499	1.0000	
	70	0.4416	0.6250	0.8334	1.0000
	100	0.2428	0.3437	0.4583	0.5499
Gill 29$^+$	10	1.0000			
	30	0.8261	1.0000		
	50	0.5046	0.6108	1.0000	
	70	0.3974	0.4811	0.7876	1.0000
	100	0.3312	0.4009	0.6564	0.8334
Shreyas Iyer 39$^+$	10	1.0000			
	30	0.6860	1.0000		
	50	0.5190	0.7564	1.0000	
	70	0.2422	0.3530	0.4667	1.0000
	100	0.0692	0.1009	0.1334	0.2859
Pandia 59$^+$	10	1.0000			
	30	0.5863	1.0000		
	50	0.3115	0.5312	1.0000	
	70	0.2266	0.3865	0.7275	1.0000
	100	0.0000	0.0000	0.0000	0.0000
Ishan 17th	10	1.0000			
	30	0.7272	1.0000		
	50	0.6364	0.8750	1.0000	
	70	0.3636	0.5000	0.5714	1.0000
	100	0.0909	0.1250	0.1428	0.2499
Jadeja 122$^+$	10	1.0000			
	30	0.4435	1.0000		
	50	0.3290	0.7418	1.0000	
	70	0.2485	0.5603	0.7552	1.0000
	100	0.0000	0.0000	0.0000	0.0000
Suryakumar 24$^+$	10	1.0000			
	30	0.6000	1.0000		
	50	0.2133	0.3555	1.0000	
	70				
	100				

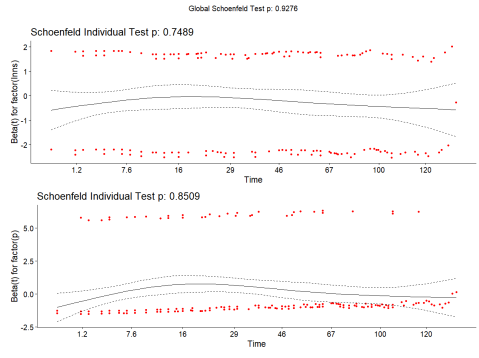


Figure 4 – PH Assumption for Kohli.

Figure 4.

For each batsman, the model includes the variable "innings", representing "batting-first" or "chasing" status ("batting-first" as baseline) and the position in which he comes to bat as covariates, while the number of runs scored is taken as the time to event, the event being "getting out". The results of the seven models, one for each batsman, with the two covariates are presented in Table 8. The PH assumption for each batsman has been verified separately using Schoenfeld residual plot and as an instance the one used for batsman Kohli is presented in Figure 4. In this Figure, the PH assumptions for the two covariates are verified by plotting the residuals against the "runs scored" and in both the plots the residuals show no pattern with respect to the runs scored ($\chi^2 = 0.1025, p > 0.05$). This enables us to conclude that the PH assumptions are satisfied for batsman Kohli. The same process is repeated for other batsmen and the PH assumptions for all batsmen are found to be satisfied.

In Table 8, each row represents the results from the Cox PH model of the individual batsman. That is, Model 1 corresponds to batsman Kohli, Model 2 corresponds to batsman Rohit and so on. From this model, it is found that innings has a significant role in getting runs for Kohli at 5% level and this is not the case for other selected batsmen in the study. Going by the fact that the corresponding estimated Hazard Ratio 0.767 is less than unity, it is observed that Kohli scores more runs when he is in "chasing" rather than "batting-first". Further, it is noted that the risk of getting out in "chasing" is 23% less compared to that of "batting-first". Also it is noted that for other batsmen, the variable "innings" is not significant and this indicates, excluding Kohli, the rest of the considered batsmen can be flexibly used in both "batting-first" or "chasing".

"Batting order" for each batsman is treated as a categorical variable, at three levels with "Top-order", "Middle-order" and "Low-order". If a batsman has played in top-order, that is taken as baseline category and compared with middle and low-orders. If a batsman has played in middle and low orders then middle-order is taken as the baseline. From Table 8, the variable "batting order" is found to be significant for batsman Rohit

TABLE 8

Summary of Cox PH Models: a, b, and c denote the number of matches played as Top, Middle, and Low-order, respectively; the "Batting Order" is treated as categorical;@ denotes baseline category; @@ denotes player not played in that order; denotes 5% level of significance.

Model Number	Player (a, b, c)	Variable	Levels	Cox PH		PH Assumption	
				HR	p-Value	Chi-square	p-Value
1	Kohli (218, 48, 0)	Innings	Batting First@	-	-	-	-
			Chasing	0.77	0.0465*	0.1025	0.7489
		Batting Order	Top@	-	-	-	-
			Middle	1.16	0.4205	0.0353	0.8510
			Low@@	-	-	-	-
2	Rohit (167, 71, 1)	Innings	Batting First@	-	-	-	-
			Chasing	0.92	0.5606	1.6087	0.2205
		Batting Order	Top@	-	-	-	-
			Middle	1.43	0.0262*	0.0396	0.9804
			Low	0.0000003	0.9935	-	-
3	Rahul (26, 26, 0)	Innings	Batting First@	-	-	-	-
			Chasing	1.20	0.5555	1.2380	0.2658
		Batting Order	Top@	-	-	-	-
			Middle	0.80	0.4910	0.1630	0.6863
			Low@@	-	-	-	-
4	Shreyas Iyer (9, 30, 0)	Innings	Batting First@	-	-	-	-
			Chasing	0.74	0.4020	0.0700	0.7913
		Batting Order	Top@	-	-	-	-
			Middle	1.58	0.2430	3.6500	0.0560
			Low@@	-	-	-	-
5	Pandya (0, 56, 3)	Innings	Batting First@	-	-	-	-
			Chasing	0.78	0.3810	3.9800	0.0460
		Batting Order	Top@@	-	-	-	-
			Middle@	-	-	-	-
			Low	1.20	0.7670	1.3200	0.2505
6	Ishan (10, 7, 0)	Innings	Batting First@	-	-	-	-
			Chasing	2.56	0.1160	0.1170	0.7323
		Batting Order	Top@	-	-	-	-
			Middle	2.53	0.1360	0.4280	0.5131
			Low@@	-	-	-	-
7	Jadeja (0, 102, 20)	Innings	Batting First@	-	-	-	-
			Chasing	1.05	0.8300	0.4060	0.5239
		Batting Order	Top@@	-	-	-	-
			Middle@	-	-	-	-
			Low	1.04	0.9040	0.8610	0.3535

at 5% level ($HR = 1.43, p < 0.05$). For rest of the batsmen, "batting order" is not significant. The Hazard Ratio for Rohit in the middle-order is 1.43 and this indicates that the risk of getting out for Rohith in the middle-order is 43% more compared to that of playing in the "top-order". That is, Rohith is expected to score more runs when he plays in the top-order rather than playing in the middle-order. This justifies his place in the opening slot, irrespective of the team "batting-first" or "chasing". It is also observed that the rest of the considered batsmen can be used in top-order or in the middle-order, leading to no significant difference in their run scoring patterns.

A single Cox PH model was developed to compare the batting performance of the batsmen, in comparison with that of player Kohli. That is, in this model, Kohli is taken as the baseline for comparison. The estimated hazard ratios of each batsman are presented in Table 9. The hazard ratios of Pandia, Jadeja and Suryakumar, in comparison

TABLE 9
*Comparison of Batsmen's performance using Hazard ratios (** denotes 1% level of significance).*

Player	Coefficient	HR	p - value
Rohit	0.1377	1.15	0.157
Rahul	0.2451	1.28	0.139
Gill	-0.0858	0.92	0.689
Shreyas Iyer	0.2339	1.26	0.194
Pandia	0.4696	1.60	$< 0.001 **$
Ishan	0.1408	1.15	0.592
Jadeja	0.4601	1.58	$< 0.001 **$
Suryakumar	0.7626	2.14	$< 0.001 **$

with Kohli are statistically significant at 1% level. The hazard ratio for Pandia is 1.599, indicating that he gets out 1.599 times more than that of Kohli. This also implies that Kohli scores more runs compared to Pandia. Similar is the case with Jadeja (HR=1.61) and Suryakumar (HR=2.142). The other batsmen do not differ significantly ($p > 0.05$) in their hazard ratios against Kohli, indicating that their performances are near equal to that of Kohli.

4.3. Parametric model

This section involves using Accelerated Failure Time parametric models with Exponential, Weibull, Log-Normal and Log-Logistic distributions for the runs scored by selected batsmen. These parametric models use a dichotomous covariate "innings", a categorical variable with "batting-first" as base line category and "chasing" as the second one. In order to address the singularities at score zero, a small positive value 0.001 is replaced

against zero and this does not alter the estimates of model parameters considerably. AIC and BIC criteria are used for identification of distributions of best-fit. A separate model is fit for each batsman. The summary results of these parametric models are presented in Table 10.

The time ratio turns out to be statistically significant at 5% level for player Kohli. For other players it is not statistically significant. For Kohli, the estimated time ratio is 1.29 and this implies that Kohli scores 29% more runs when he chases compared to "batting-first". The distribution of best-fit, using the criteria of AIC and BIC concludes that Weibull distribution fits well for Kohli, Rohith, Rahul, Pandia, Jadeja and Suryakumar. For Gill and Ishan, Exponential distribution fits well. In case of Shreyas Iyer, AIC suggests Weibull distribution, whereas BIC suggests Exponential distribution. Thus, it is observed that in all but one case, both AIC and BIC indicate the same distributions and in most of the cases it is Weibull distribution.

5. Discussion and Conclusion

Considering the number of runs scored by a batsman as his survival time and with the information on whether "batting-first" or "chasing" and the position in which he is made to bat are used in the analysis of this study. Survival probability estimates, Hazard ratios, Time ratios and conditional probabilities derived from this information are used to measure and predict the performance of selected Indian batsmen. This study helps to find out the performance metrics of the batsman that will help us placing him in a particular batting slot for "batting-first" or "chasing" in the ODIs.

Using Kaplan-Meier method, it is observed that there is no significant difference in scoring runs between "batting-first" and "chasing" for all players except Kohli. This is also reiterated using the Hazard ratio in the Cox PH Model. The batmen's performance, measured in terms of scoring 50's and centuries were estimated using Kaplan-Meier survival probabilities. The survival probability estimates further indicate that there is a significant difference in the run scoring pattern between the top-order and the middle-order batsmen, in that it is seen top-order batsmen score comparatively higher number of runs. This may be due to the hidden factor of "field restrictions", that go in favor of the top-order batsmen. But, when compared among the top-order batsmen, the study observes no significant difference in their run scoring pattern, whereas significant difference is seen among the middle-order batsmen. Conditional probabilities estimated using Kaplan-Meier method help the team management to suggest a proper sequence of batsmen to be used in a match and this optimizes the chances of getting more runs for the players and as well for the team. Further, going by these metrics against selected countries, the study also suggests best players combination against those countries. The position of the batsmen, after being classified as top, middle and low orders, does not significantly alter their run scoring performance of all batsmen except Rohith. For Rohith, the Cox PH model reveals that he tends to score more runs while playing in the top-order compared to being placed in the middle-order. The Cox Model also helps

TABLE 10
Parametric survival models with covariate "Innings".

Player		Exponential	Weibull	Log-Normal	Log-Logistic
Kohli	Time Ratio	1.29	1.47	2.69	1.53
	p-Value	0.054	0.060	0.008	0.102
	AIC	2280.48	2220.35	2365.16	2286.99
	BIC	2287.64	2231.10	2375.91	2297.74
Rohit	Time Ratio	1.04	1.25	1.81	1.40
	p-Value	0.766	0.340	0.139	0.245
	AIC	1996.74	1912.49	2033.65	1965.79
	BIC	2003.70	1922.92	2044.08	1976.22
Rahul	Time Ratio	0.80	0.78	0.33	0.56
	p-Value	0.458	0.536	0.114	0.251
	AIC	426.71	422.37	447.55	431.89
	BIC	430.61	428.22	453.40	437.74
Gill	Time Ratio	0.66	0.70	1.45	0.77
	p-Value	0.326	0.461	0.693	0.657
	AIC	250.01	251.27	270.07	257.80
	BIC	252.74	255.37	274.17	261.90
Shreyas Iyer	Time Ratio	1.41	1.31	1.42	1.46
	p-Value	0.332	0.323	0.327	0.268
	AIC	350.21	349.10	356.39	356.24
	BIC	353.54	354.09	361.38	361.23
Pandia	Time Ratio	1.29	1.36	1.43	1.73
	p-Value	0.366	0.423	0.653	0.264
	AIC	465.97	461.23	504.78	480.76
	BIC	470.13	467.46	511.01	486.99
Ishan	Time Ratio	0.44	0.44	0.36	0.36
	p-Value	0.102	0.107	0.108	0.117
	AIC	157.53	159.52	161.60	162.25
	BIC	159.19	162.02	164.10	164.75
Jadeja	Time Ratio	0.92	0.91	1.27	1.02
	p-Value	0.727	0.745	0.681	0.962
	AIC	712.31	703.81	761.02	722.76
	BIC	717.92	712.22	769.43	731.17
Suryakumar	Time Ratio	1.31	1.45	0.59	1.20
	p-Value	0.549	0.646	0.730	0.871
	AIC	179.69	170.79	184.59	179.58
	BIC	182.05	174.32	188.13	183.11

us compare batsmen, based on their estimated hazard ratios and their statistical significance. The results of AFT parametric models help us to rank the batsmen in terms of their run scoring pattern, quantified by the values of the acceleration factor and their statistical significance.

The results from the study will help the team management not only to select best players against a specific country, but also the order in which they should be sent in to bat. Further, the given the information on "batting-first" or "chasing", the management can accordingly rotate the order of batting of their players. Estimates on the conditional probabilities help the captain of the team to decide about the next player to be brought in for batting at crease, particularly, during the live-match scenario.

All such estimates and recommendations could be given,simply based on the average runs scored by each batsman in different categories of their play. But, the added advantage of this study lies in using survival models that takes in to account the incomplete information, also known as censoring, to the maximum possible extent.

5.1. Limitations of the study

Our study does not use information on relevant covariates, such as, nature of the pitch, time of play (Day or Day-Night), opposition team's strength in terms of phase and spin bowling, position of the series (already won/lost/tied) and the place of the series (Home Ground/Away) and this might have reduced the accuracy of our estimates and predictions. This study includes batsmen observed from a narrow time window and this affects the estimates of the performance metrics of the players.

5.2. Future work

The Cox model in this study could be improved by adding more control variables listed in the limitations of the study and this will improve the sensitivity of the results. A broader time window for each batsman would serve to get us more stable results. Further, a more realistic time-dependent Cox PH model could be used for the future work. As, by now we have the results of Men's World cup 2023, our predictions could be verified against the real time data.

Acknowledgements

The authors submit their sincere gratitude to the reviewers for their comprehensive and critical comments and suggestions, without which the final version of this paper would be a distant reality.

References

D. Cox (1972). *Regression models and life tables.* Journal of the Royal Statistical Society, Series B, 34, pp. 187–220.

B. J. KACHOYAN, M. WEST (2018). *Deriving an exact batting survival function in cricket.* 14th Australasian Conference on Mathematics and Computers in Sports, pp. 62–67.

P. KALPDRUM, P. NIRAV KUMAR (2018). *Increased prediction accuracy in the game of cricket using machine learning*. International Journal of Data Mining & Knowledge Management Process (IJDKP), 8, no. 2, pp. 19–36.

E. KAPLAN, P. MEIER (1958). *Nonparametric estimation from incomplete observations*. Journal of American Statistical Association, 53, pp. 457–481.

N. P. MANOJ ISHI, JAYANTRAO PATIL, V. PATIL (2022). *Winner prediction in one day international cricket matches using machine learning framework: An ensemble approach*. Indian Journal of Computer Science and Engineering, 13, no. 3, pp. 628–641.

N. MOHAN, M. RAMAKRISHNAN, R. RAVANAN (2022). *The comparison of survival approach with fuzzy logic for national stock exchange data during covid-19 in India*. International Journal of Mathematics and Statistics, 23, no. 2, pp. 41–51.

H. PREETHAM, R. PRAJWAL, P. KUMAR, N. KUMAR (2023). *Cricket score prediction using machine learning*. International Journal of Innovative Research in Technology, 9, no. 8, pp. 109–114.

P. SHAH, R. CHAUDHARI, P. M.N. (2023). *Evaluating batsman using survival analysis*. Statistics and Applications, 21, no. 1, pp. 51–62.

P. SHAH, M. PATEL (2018). *Ranking the cricket captains using principal component analysis*. International Journal of Physiology, Nutrition and Physical Education, 2, no. 3, pp. 477–483.

O. STEVENSON, B. BREWER (2017). *Bayesian survival analysis of batsmen in test cricket*. Journal of Quantitative Analysis in Sports, 13, no. 1, pp. 25–36.