STAT 421: Project Work

Biostatistician's Insight into Sport Analysis

Submitted by

Soumarya Basak

C91/STS/211031

Submitted on

09 June 2023

Guides

Prof. Bhaswati Ganguli

Prof. Gaurangadeb Chattopadhyay

Prof. Sugata Sen Roy

Abstract

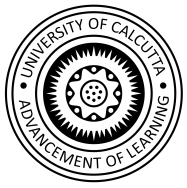
In this project we will demonstrate the application of survival analysis beyond biostatistics in the field of sports . Considering the upcoming ICC tournaments in this year we will study the performance of some top order batsmen to predict their performance in those tournaments based on their past performances. Based on few criteria we have picked five such batsman from two teams India and Australia.

We consider the data as a *censored data*, the most important for survival analysis. The event is then — Out in an innings. That is if a batsman is out in an innings then there is an event, if he remains not out it is a censored observation. The project will include testing equality of two survival probabilities at the crucial point of a innings for the batsmen. Additionally, regression-type models will be employed to gain further insights into their survival rates and expected run scores in the tournaments.

Acknowledgement

I would like to express my heartfelt gratitude to my project guides, *Prof.* Bhaswati Ganguli, *Prof.* Sugata Sen Roy, and *Prof.* Gaurangadeb Chattopadhyay, for their invaluable guidance and for providing me with the opportunity to explore my thoughts in this project. Their expertise and support have been instrumental in shaping the direction of my work. I would also like to extend my thanks to the other professors in the Department of Statistics, University of Calcutta, *Prof.* Rahul Bhattacharya, *Prof.* Arindam Sengupta, and *Prof.* Asis Kumar Chattopadhyay, for their guidance and contributions throughout my master's journey along with my guides. Their insights and encouragement have played a significant role in the development of my thoughts and ideas. I am deeply grateful to the Department of Statistics, University of Calcutta, for providing me with a conducive environment to learn and grow. Their resources and academic rigor have greatly contributed to my overall learning experience. Last but not least, I would like to thank my family and friends for their unwavering support and encouragement. Their presence and discussions have been a source of inspiration and motivation throughout this journey.

I am truly honored to have been a part of this enriching experience, and I will cherish the knowledge and guidance I have received from all those mentioned above.



Department of Statistics
University of Calcutta

Contents

1 | Objective

The year 2023 promises to be an exhilarating one for cricket enthusiasts, as they gear up for two prestigious tournaments: the 2nd edition of the *World Test Championship final* set to take place from 7th to 11th June at the iconic venue of The Oval in London, England. Following this thrilling contest, the cricketing world's attention will shift to India, as the 13th edition of the *ICC Cricket Men's World Cup* unfolds between 5th October and 19th November.

Both India and Australia will be among the teams participating in these prestigious tournaments, adding an extra layer of excitement to the proceedings. As the anticipation builds, every team desires a strong start to their matches in these crucial tournaments. In this analysis, we will delve into the performances of the top-order batsmen from both India and Australia, who hold the key to their teams' success on the grand stage.

Based on the ICC ranking of the batsman in both in ODI and test from October 2022 to January 2023 and also played sufficient no. of matches, we have picked 5 players from the two teams and will try to compare their performance in those tournaments. We choose these 5 players except Steve Smith and Virat Kohli as they are already somewhat better than others.

In this analysis, our objective is to evaluate the performance of KL Rahul, Rohit Sharma, Shikhar Dhawan, David Warner, and Marnus Labuschagne leading up to the Test Championship and World Cup. We will assess which openers have the potential to secure a spot in the playing 11 for India and investigate whether there are performance differences among the top order batsmen from the two teams. Additionally, we will explore if the players' performances vary based on the format of the game (ODI or Test), the venue of the match, and the type of innings (chasing or defending). Moreover, we will examine the risk of being dismissed at different run thresholds and analyze which types of dismissals pose a greater challenge for the batsmen. By considering these aspects, we aim to gain insights into the strengths and weaknesses of the selected batsmen and understand their potential impact on the team's performance in the upcoming tournaments.

2 | Data

2.1 | Collection

We possess comprehensive data on the international cricket careers of KL Rahul, Shikhar Dhawan, Rohit Sharma, David Warner, and Marnus Labuschagne, covering their performances across all three formats (Test, ODI, and T20I) from their debut until January 15, 2023. With this wealth of information, we can provide valuable insights into the performance of these players, shedding light on their achievements and contributions to their respective teams.
The data for the 5 players is been collected from statguru at espncricinfo.com.

2.2 | Description

The raw data for a player consist of 13 columns. For the each of the player there are same variables in the datasets. First 10 rows of KL Rahul's data is shown below,

Runs	Mins	BF	4s	6s	SR	Pos	Dismissal	Inns	Opposition	Ground	Start Date	Match no
3	12	8	0	0	37.5	6	caught	2	Test v Australia	Melbourne	26-Dec-14	Test # 2152
1	6	5	0	0	20	3	caught	4	Test v Australia	Melbourne	26-Dec-14	Test # 2152
110	356	262	13	1	41.98	2	caught	2	Test v Australia	Sydney	06-Jan-15	Test # 2156
16	56	40	3	0	40	2	caught	4	Test v Australia	Sydney	06-Jan-15	Test # 2156
7	13	7	0	0	100	1	lbw	2	Test v Sri Lanka	Galle	12-Aug-15	Test # 2176
5	19	14	0	0	35.71	1	lbw	4	Test v Sri Lanka	Galle	12-Aug-15	Test # 2176
108	268	190	13	1	56.84	2	caught	1	Test v Sri Lanka	Colombo (PSS)	20-Aug-15	Test # 2177
2	4	3	0	0	66.66	2	bowled	3	Test v Sri Lanka	Colombo (PSS)	20-Aug-15	Test # 2177
2	1	2	0	0	100	1	bowled	1	Test v Sri Lanka	Colombo (SSC)	28-Aug-15	Test # 2179
2	16	8	0	0	25	2	bowled	3	Test v Sri Lanka	Colombo (SSC)	28-Aug-15	Test # 2179

Where the columns represents:

- `Runs:` run that the batsman made in the innings.
- `Mins:` How many minutes the batsman spent in the crease.
- `BF:` No. of ball faced.
- `4s:` No. 4s the batsman made.
- `6s:` No. of 6s the batsman made.
- `SR:` Strike rate of the batsman.
- `Pos:` In which position he batted.
- `Dismissal:` How as he dismiss.
- `Inns:` The innings of the match when he batted.
- `Opposition:` The type of the match with the opponent's name.
- `Ground:` In which ground the match was played.
- `Start Date:` The date of the match.
- `Match no.:` No. of the match internationally for the corresponding format.

2.3 | Preprocessing

Before delving into a detailed analysis, it is crucial to preprocess the data to ensure its suitability for the upcoming analysis. This involves performing necessary transformations on certain columns, which is known as Data Preprocessing. The following steps will be taken to make the data more workable and conducive for analysis. Keep in mind that all the preprocessing mentioned here is performed for all the 5 data sets separately.

Subseting the data sets

Our raw data consist of the information of three formats played by the player internationally. But for our purpose we only use the information of *Test* and *ODI* format as *T20I* is an unstable format, the batters has to bat rather in a different manner than ODI and Test. So to make the analysis unaffected by this, we drop the information of T20Is. And further more manipulation is done on the subsetted datasets.

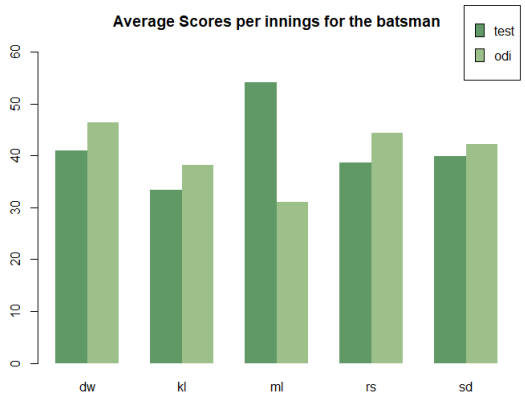
- `Runs`: The `Runs` column is the most crucial column for our overall analysis, but it was not well defined, i.e. there are some character values as `DNB, TDNB`, that make the overall variable character, so we take care of those values and convert it into numeric.
- `delta`: We created an indicator variable `delta` which is
  1: if the batsman got out in the innings and
  0: if he remains not out.
  It is basically an indicator that indicates 1 if event occur and 0 is the observation is censored. Here we are considering event if a player got out in an innings, and censor if he remains not out.
- `chasing`: Created a variable, `chasing` which is categorical in nature with 2 levels
  `d`: the batsman scores the innings while defending that innings and
  `c`: the batsman score the innings while chasing that innings.
- `Venue`: We also created a variable `Venue` which is
  `H`: if the batsman played that innings in home ground,
  `A`: if the batsman played the innings in away ground.
- `Year`: We created column `Year` by extracting the year of the match from the `start Date` column, as we are planning for a yearly analysis.
- `Type`: We separated the type of the match from '`Opposition`' column and store in the column.
- `status`: We created a column `status` indicates
  0: If the batsman is not out or run out. (censor observation)
  1: When the batsman gets dismissed by being bowled.(event due to cause 1)
  2: When the batsman gets dismissed by being caught.(event due to cause 2)
  3: When the batsman gets dismissed by being lbw.(event due to cause 3)

3 | Exploratory Analysis

Before delving deeper into our analysis, it is essential to explore some key features that will provide us with valuable insights.

N.B. Please take note that throughout the analysis, we have utilized the average runs of the batsmen in terms of “average runs per innings" rather than “average runs per out." Since we are considering the censored data so, the usual average is bit questionable, so avoid such problem we remove the information of censoring by choosing average runs per innings which may be an underestimate of our desire quantity.

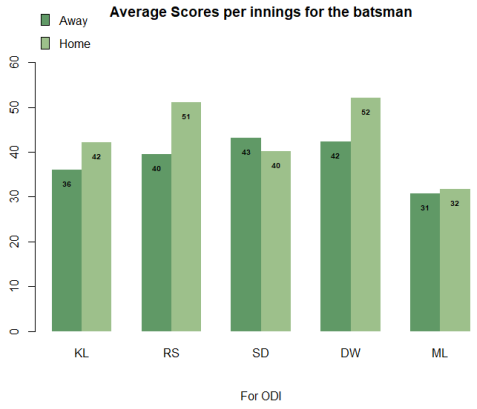
3.1 | Batting Averages of the players

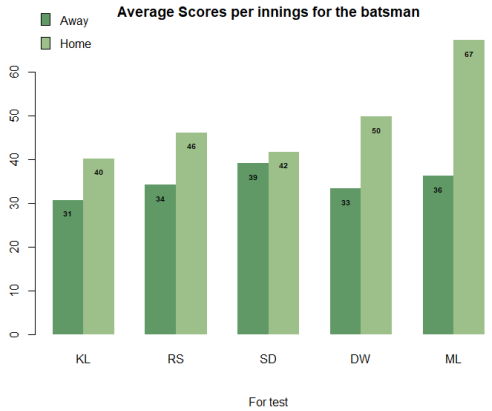


We possess the batting averages of the players for different cricket formats since 2013. Upon initial observation, we note that all players, except Marnus Labuschagne, have better averages in ODI cricket. Labuschagne stands out with a test average of 54.1 runs per innings, highlighting his importance in Australia’s test team.

Comparing KL Rahul and Shikhar Dhawan, their averages are relatively similar, with Dhawan having a slightly better average of 39.1 in test 42.2 in ODI, those for Rahul is 33.3 and 38.1 respectively. Additionally, David Warner’s presence provides a strong composition to the team alongside Rohit Sharma. Both the average for Warner is above 40, whereas for Rohit it is 38.7 in test and 44.4 in ODI, both lower than Warner’s.
However, it is important to acknowledge that these observations are based on aggregated data and there may be underlying factors that require a more in-depth analysis.

3.2 | Batting Averages of players for different situation





Let's start by discussing Marnus Labuschagne's batting averages, which reveal some intriguing figures. While his ODI average hovers just above 30 in both home and away conditions, his test average sees a significant surge, almost doubling to an impressive 67.14 in home conditions. This indicates his formidable presence and effectiveness when playing on his home grounds. However, in away situations, Labuschagne's batting average drops to 36, providing some respite for opposing teams. It suggests that he may not be as dominant or impactful when playing outside his familiar surroundings. Upon examining the visual representations, we can observe that in the ODI format, both Rohit Sharma and David Warner have impressive batting averages of over 51 runs per innings. This places them ahead of KL Rahul and Shikhar Dhawan, who have averages just above 40 in their home conditions. In away conditions, Shikhar Dhawan stands out as the most impactful batsman with an average of 42.8 runs per innings. He is closely followed by David Warner, Rohit Sharma, and KL Rahul, who have averages of 42, 40.2, and 36.3, respectively, in such situations.
Considering the upcoming *World Cup*, it is important to note that the playing conditions will be in favor of Rohit Sharma, KL Rahul, and Shikhar Dhawan as they will be playing on home turf. On the other hand, David Warner and Marnus Labuschagne will be competing away from their familiar conditions.
Based on these parameters, Rohit Sharma emerges as a key player to watch from the Indian team, given his ability to perform exceptionally well in home conditions. Similarly, David Warner is expected to be a strong competitor for KL Rahul, providing tough competition from the Australian side, both with average close to 42.

In the Test format, we observe a significant drop in batting averages for all Indian batsmen in away conditions compared to their home performances. Similarly, David Warner also experiences a decline in his average, managing 33.3 in away situations, whereas he had an average of above 50 in home conditions.

Considering the upcoming *World Test Championship final*, where the batters will face an away situation, we can expect their average runs to range between 30 to 35, with the exception of Shikhar Dhawan. However, another crucial factor to consider is the amount of time spent on the crease, which can have a significant impact on the game.

When examining the bars representing away situations, we find that Rahul spends the most time on the crease, averaging 95.5 minutes. Labuschagne closely follows him, staying at the crease for an average of 94.8 minutes, causing havoc for bowlers. Rohit also contributes to stabilizing the lineup, spending around one and a half hour at the crease. On the other hand, David Warner

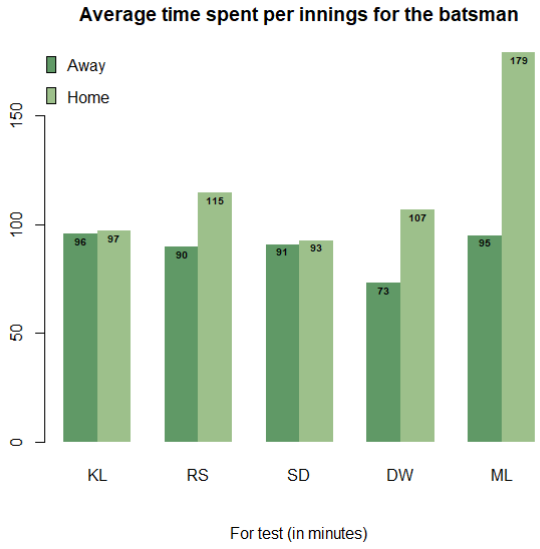


Figure 3.1: Represents the average time spent by the batsmen in the crease in Test matches

has a relatively shorter average crease time of 75 minutes, indicating a slightly more vulnerable position compared to the others.

However, batting averages alone do not always provide a complete picture of a player's performance, especially considering that these averages are calculated from their debut until January 2023. Over such a long period, players may experience fluctuations in their performance, which can influence their selection for future games. Therefore, it becomes crucial to analyze how their performance has changed over the years to gain a better understanding of their overall trajectory.

Let's delve into a detailed examination of how these players' performances have evolved throughout their careers.

3.3 | Yearly Change in Batting Average

When examining the year-by-year performance of these players from their debut, it is evident that their average scores per innings have seen various ups and downs. Shikhar Dhawan's international test appearances have been limited since 2018, resulting in a completion of that year with an average of 27.4. KL Rahul, on the other hand, witnessed a surge in his average, which subsequently dropped in 2022 from 46.1 to 17.1. This decline in form has opened up an opportunity for Shubman Gill to potentially fill that position.

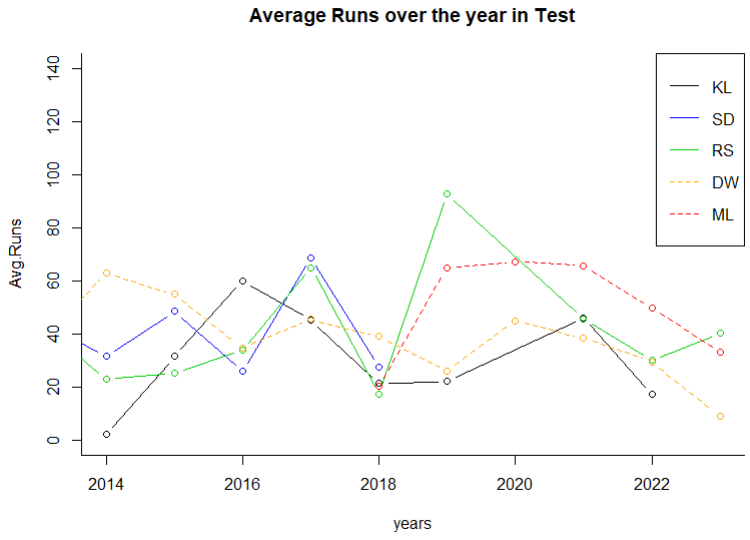
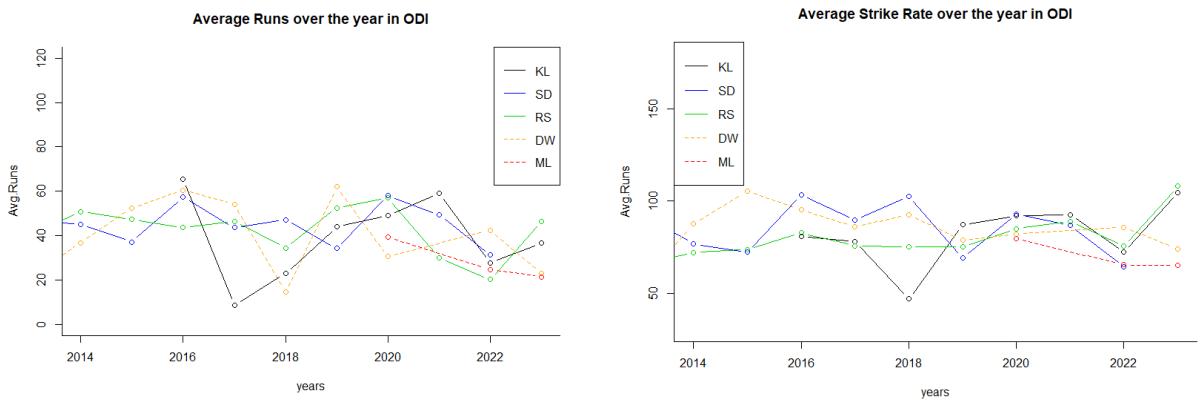


Figure 3.2: Batting averages of the batsmen from 2014 to 2023 in Test format

In contrast, Rohit Sharma and Marnus Labuschagne have shown a gradual upward trend in their averages over time. Rohit's average has improved from the 20s during 2014-15 to a commendable 40.3 in 2023, with a solid performance in 2022, recording an average above 30. Labuschagne started his debut year with an average of 20.4, despite limited innings, and showcased his potential by maintaining an average above 65 in the subsequent years. However, his average dipped to 49.9 in 2022.

In contrast, David Warner, who had impressive averages of 63 and 54 in 2014-15, has witnessed a decline in recent years, with averages of 38 and 29 in 2022 and 2023, respectively. This suggests a clear downward trend in his performance in the test format.



(a) Average Runs

(b) Average strike rates

Figure 3.3: Batting averages with the average Strike rate of the batsmen from 2014 to 2023 in ODI format

When analyzing the ODI format, we have observed the changes in strike rates and average run scores over the years. KL Rahul's average run score has shown a clear upward trend from 2017 to 2023, culminating in an average of 36.7. Notably, he achieved a high average score of 65.3 in 2016, which solidified his position in the team. Furthermore, his strike rate has also increased over the years and currently stands at 104.0. It's worth mentioning that Rahul has been given a middle-order spot in the team since the last quarter of 2022.

Shikhar Dhawan, on the other hand, has maintained a consistent batting average of over 40 throughout the years. However, his strike rate has experienced a downward trajectory since 2016, currently resting at 64.3 in 2023. A similar case can be observed for David Warner, although his average score has remained relatively low in recent years.

Turning our attention to Rohit Sharma, we see an upward trend in his average score, consistently ranging between 40 and 50 over the years. This indicates his consistent performance in the ODI format.

In contrast, Marnus Labuschagne's performance in ODIs has been less impressive, particularly in his debut year. His average score has not been noteworthy compared to the other players discussed.

Impact In the tournaments:

After examining the visualizations, we can identify Labuschagne as the standout player to watch in the World Test Championship final. His impressive batting average and extended time spent at the crease set him apart, particularly when compared to Rahul, who possesses a similar average. Furthermore, Rohit's recent performance in the format surpasses that of Rahul, making him another player to keep an eye on.

Shifting our focus to the World Cup, Rohit emerges as a formidable threat with an average of 46.4 in home conditions and a striking rate of 108.0. His current form indicates his potential to dominate in ODI cricket. Additionally, Rahul's recent outstanding performances position him as the second most dangerous player on the list.

4 | Survival Analysis

Having gained insights from our exploratory analysis regarding the average runs per innings, we now shift our focus to the batsmen's survival on the crease. The average runs scored by a batsman is highly influenced by the duration of their stay at the crease. Therefore, we will delve deeper into their survival patterns and analyze the factors contributing to their longevity in an innings.

4.1 | Idea of Survival Analysis

Let T be a time to occur an event. Hence random and $T > 0$. Assume,

$$T \sim F(t), \quad \text{with density } f(t), \quad t > 0$$

Then the *survival function* is defined as,

$$S(t) = P[T > t] - 1 - F(t) \tag{4.1}$$

The *hazard function or the instantaneous failure rate* is defined as,

$$h(t) = lim_{\Delta t \to 0} \frac{1}{\Delta t} P[t \leq T \leq t + \Delta t | T > t]$$

$$= \frac{f(t)}{S(t)} = -\frac{d}{dt} logS(t) \tag{4.2}$$

And the cumulative function is defined as,

$$H(t) = \int_0^t h(u)du = -logS(t) \tag{4.3}$$

Based on these three functions we will perform our analysis. So based on sample we need to obtain an estimate of these functions.

4.1.1 | Kaplan-Meier Estimation of Survival Function

The Kaplan-Meier estimator is a widely used method for estimating survival curves. It does not rely on any distributional assumptions about the event time variable T, making it a non-parametric approach.

Suppose our observation times are: $y_1 < y_2 < ..... < y_n$ (wlg).
Then divide the study period in to several intervals with end points $(y_{i-1}, y_i]$ for $i = 1(1)n,\ y_0 = 0$.
Let,
d_k : # event occur at y_k
C_k : # censoring occur at y_k
n_k : # individuals at risk at y_k

Now, define $p_k = P[$ an individual survives $y_k |$ the individual has survived $y_{k-1}]$

$$= P[T > y_k | T > y_{k-1}] \ , \quad j = 2, 3...$$

and $p_1 = P[T > y_1]$
Also define $q_k = 1 - p_k$

Then $s(y_j) = P[T > y_j] = \prod_{k=1}^{j} p_k$
Let,

$$\delta_i = \begin{cases} 1 & if \quad \text{ith observation has event} \\ o & if \quad \text{ith observation is censored} \end{cases} \quad , i = 1, 2, ...n$$

be the censoring indicator.
Assuming that all the censoring take place after the events, the estimates of q_k and p_k are,

$$\widehat{q_k} = \begin{cases} 0 & if \quad \delta_k = 0 \\ \frac{d_k}{n_k} & if \quad \delta_k = 1 \end{cases} \qquad \& \quad \widehat{p_k} = \begin{cases} 1 & if \quad \delta_k = 0 \\ 1 - \frac{d_k}{n_k} & if \quad \delta_k = 1 \end{cases}$$

Then the *Kaplan Meier Estimator* of survival function is given by,

$$\widehat{S(t)} = \prod_{j:y_j \leq t} \widehat{p_k} = \prod_{j:y_j \leq t} (1 - \frac{d_k}{n_k})^{\delta_k} \tag{4.4}$$

Since it is an estimate, we need some precision upon this. The estimator of variance of the estimate is given by,

$$\widehat{V(\widehat{S(t)})} = [\widehat{S(t)}]^2 \sum_{j:y_j \leq t} \frac{d_j}{n_j(n_j - d_j)} \tag{4.5}$$

Equation 4.5 is known as *Greenwood's formula.*

Using equation(4.4) we have estimates of *cumulative hazard function* and *Hazard function* as,

$$\widehat{H(t)} = -log\widehat{S(t)} \tag{4.6}$$

$$\widehat{h(t)} = \widehat{H(t)} - \widehat{H(t-)} \tag{4.7}$$

4.1.2 | Mantel Haenszel Test

Mantel Haenszel test, also known as *Log Rank test* is a statistical test to test the equality of two or more survival functions. For our analysis we will only use this to test the equality of two survival functions so we will discuss this for the case.

Let for Group 1,
Event time $= T_{1i} \sim S_1(t)$, censoring time $= C_{1i} \sim G_1(t)$ for $i = 1(1)n$
where we observe $Y_{1i} = min(T_{1i}, C_{1i})$ with event indicator δ_{1i}

For Group 2,
Event time $= T_{21i} \sim S_2(t)$, censoring time $= C_{2i} \sim G_2(t)$ for $i = 1(1)n$
where we observe $Y_{2i} = min(T_{2i}, C_{2i})$ with event indicator δ_{2i}

Assume that $G_1(t) = G_2(t) \forall t$
To test:

$$H_0 : S_1(t) = S_2(t) \ \forall t \quad ag \quad H_1 : S_1(t) \neq S_2(t) \tag{4.8}$$

The test can be perform by the following steps:

- Combine the two sets of data and order the observations.
- Consider each uncensored time points say τ_j and conduct the following 2×2 table,

$T = \tau_j$	E	NE	
G_1	d_{ij}		n_{1j}
G_2	d_{2j}		n_{2j}
	m_{1j}	m_{2j}	n_j

$n_{ij} + n_{2j} = n_j$
n_{kj} =# of times at risk at τ_j for kth group, k=1,2
d_{kj} =# events at τ_j for kth group, k=1,2

Define,

$$O_j = d_{1j} \ , \quad E_j = E(d_{1j}) = \frac{n_{1j}m_{1j}}{n_j} \ , \quad V_j = \frac{m_{1j}m_{2j}n_{1j}n_{2j}}{n_j^2(n_j - 1)}$$

Also define,

$$O = \sum_j O_j \ , \quad E = \sum_j E_j \ , \quad V = \sum_j V_j$$

Now the *Mantel Haenszel* test statistic is defined as,

$$MH = \frac{(O - E)^2}{V} \longrightarrow \chi_1^2$$

4.2 | Kaplan Meier fit to the Data

To conduct an in-depth survival analysis, we use the 'Runs' column as our time variable along with the 'delta' column we have created. The 'delta' column serves as our event indicator, with a value of 1 indicating that the batsman was out in the innings and a value of 0 indicating that the batsman was not out. This approach allows us to analyze the survival probabilities and hazard rates associated with each batsman's innings.
To address the issue of sparsity in scores of the batsmen, scores exceeding 120 were censored at 120 for each of them. This approach helps in reducing the impact of limited times extreme scores and ensures that the dataset remains balanced and manageable for analysis.
Now, to have a look onto the survival probabilities of the batsman we estimated their survivals in a *Non parametric* manner[1].

One of the most widely used methods for that is the *Kaplan-Meier estimation*. A detailed explanation of this estimator is provided in Section 4.1.1.
Figure 4.1 shows the *Kaplan-Meier* fit of the survivals of the batsmen, for the two formats separately with a *confidence band* of the estimates. As the curves are close to each other we test for the equality of the survival curves for the two formats for each of them. This can be done by *Mantel Haenszel Test* discussed in section 4.1.2. The *p values* of the chi-squared test is given along with the plots of the survival curves for each of them for testing the hypothesis 4.8.

Looking at the survival curves of the batsmen we can clearly see that except Labuschagne all of them is like to survive better in ODI format. However the p values of log rank tests

[1]The survival curves hardly follows any parametric pattern. I have tried to fit weibull, log-logistic, gompartz, log-normal distribution but none them gives an overall close fit to the KM estimates. Although for a range of runs few gives a good fit.

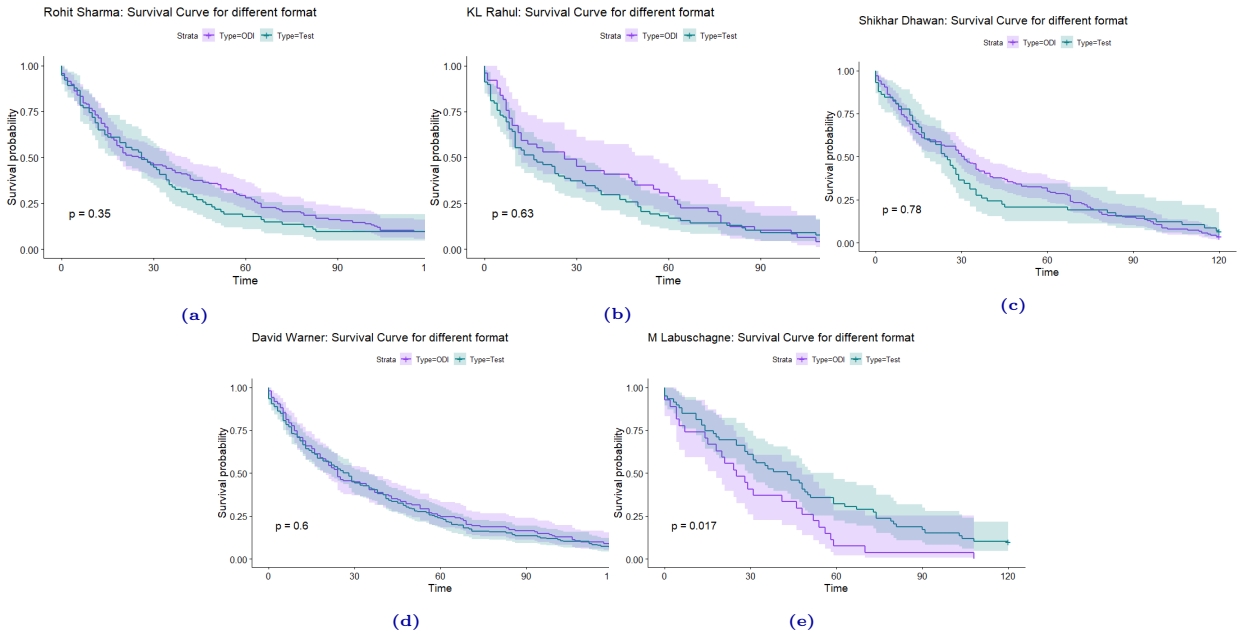


Figure 4.1: KM fit of the survival function for each of the batsmen for the two format of the cricket, with a confidence band of the estimates and *p value* of MH test

say that the difference of the survival curves is not statistically significant at 5% level. Focusing on Labuschagne, he very likely to survive in test matches than ODIs with an almost uniform decrease in survival curve whereas for the others it has drop at the early time of the innings. The p value of the log rank test 0.017 suggesting that his test performance is significantly better than ODIs at 5% level.

4.3 | Hazard curves of the batsmen

While we have gained insights about the performance in test, it is challenging to make direct comparisons for the ODI format from the survival curves, as being a decreasing curve it is more or less similar for all of them. It would be more appropriate to study their hazards. Hazard or Instantaneous failure rate may be of any shape not necessarily decreasing like survival. Using the relation in 4.7 obtain the hazard curve of the batsmen below. From the hazard curve we can see that Rahul has a higher hazard at the begin-

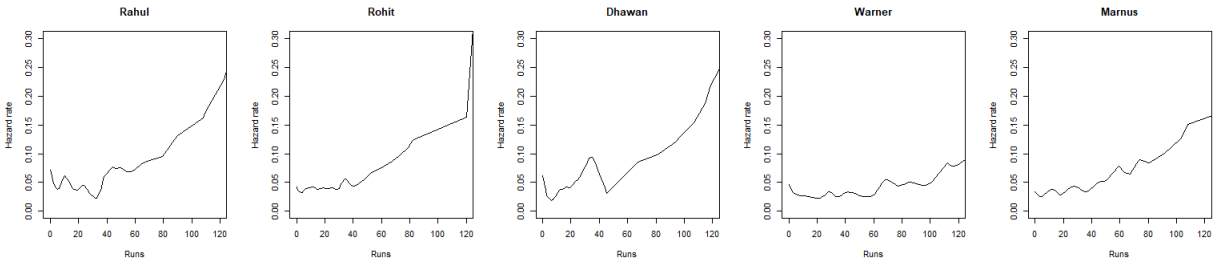


Figure 4.2: hazard curves of the batsmen for *test* format

ning of his innings than any other which is decrease later on but again increase before 40s. A similar pattern can be found for Rohit rather he is just stable than Rahul in beginning, almost as Labuschagne. Forgetting Dhawan in the comparison as he played

his last test match in 2018. Hazard of Warner is pretty better than others almost similar throughout the whole span, but form of recent years is a big issue then. Coming to the best among them, having a smaller hazard uniformly over the span over beginning to the 40s Labuschagne has a lower hazard than Rahul and Rohit.

Examining the hazards for the ODI format, we observe that Rohit and Warner have relatively constant hazards. Shikhar Dhawan closely follows them, but his hazard starts to increase after reaching the 80-run mark. On the other hand, Rahul exhibits the most instability among the batsmen, with a sudden increase in hazard between scores of 20 and 40. This suggests that once Rahul gets a start, he is more likely to be dismissed within that score range compared to the other batsmen. Furthermore, his hazard increases rapidly after reaching a score of 60. Labuschagne's performance in ODIs shows less stability compared to the other batsmen as we can see from his history of 27 ODI innings.

Thinking about the *World cup*, Rohit would be a good performer among these five

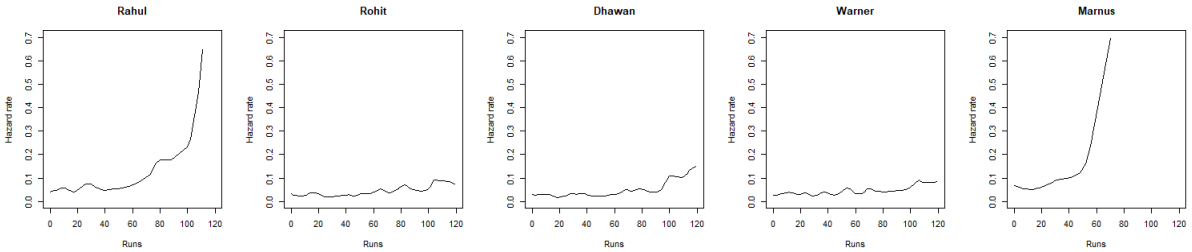
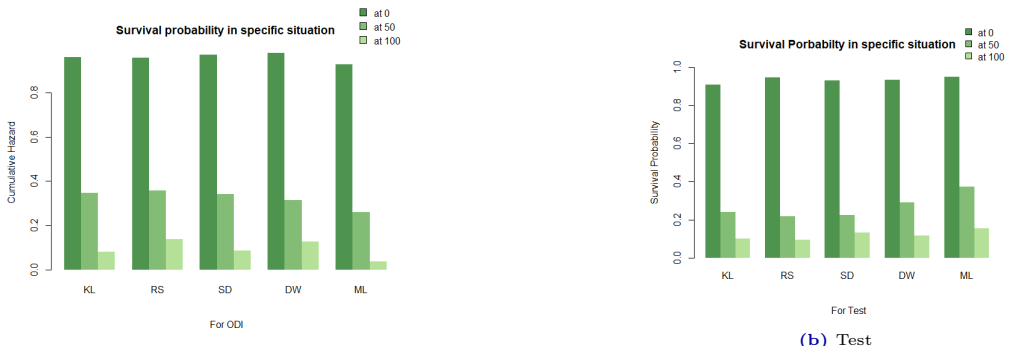


Figure 4.3: hazard curves of the batsmen for *ODI* format

considering the current year's performance. The second player would be Warner though he will not be that explosive as he used to be earlier but he is likely to score a higher runs like Rohit.

4.4 | Survival probabilities at crucial time

To gain valuable insights into the potential of batsmen to reach significant milestones, such as scoring their first run or nearing the landmarks of 50 or 100 runs, we examine their survival probabilities at these critical stages. By analyzing their survival probabilities at these specific points, we can better understand their ability to achieve these important milestones and gauge their performance accordingly.



(a) ODI

(b) Test

Figure 4.4: Plot represents the survival probabilities at 0, 50 and 100

Upon comparing the survival probabilities, it is evident that all of them have similar chances of survival at the beginning (at 0 run), with Warner potentially having a slightly higher probability. However, the differences may likely to be insignificant. As we move towards the milestone of a half-century, Rohit Sharma stands out with a higher chance compared to the others, while Labuschagne has the lowest probability. When it comes to reaching the century milestone, Rohit Sharma exhibits a high potential to score a century, surpassing the others. Rahul, Dhawan, and Warner have nearly equal chances of reaching the century mark.
Moving toward to the test format, Rahul has a lower probability of surviving at 0 whereas Labuschagne has the highest, getting a almost close competition by Rohit Sharma. A very similar result can be observed with the survival probabilities at 50. Labuschagne is the best there keeping a good difference with Rohit who has the lowest then indicating that even though he survives better in beginning his failure rate increases later before a half century in test cricket. moving forward to the century there we can hardly see any variation in the survival probabilities of the batsmen.

Now, these observations are made based on a preliminary analysis of the plots in Figure 4.4. To further validate our findings, we will conduct statistical tests to assess the equality of survival probabilities at specific points for two batsmen.

4.4.1 | Test for Significance

We aiming to test at a fix time point t_0, the equality of survival probabilities. i.e. we wish to test,

$$H_0 : S_1(t_0) = S_2(t_0) \quad vs \quad H_1 : S_1(t_0) \neq S_2(t_0)$$

Assuming the estimates are approximately normal then a large sample test can be perform base of the statistic,

$$Z = \frac{\widehat{S_1(t_0)} - \widehat{S_2(t_0)}}{\sqrt{\hat{V}(\hat{S}_1(t_0)) + \hat{V}(\hat{S}_2(t_0))}} \overset{H_0}{\sim} N(0,1)$$

where, $\widehat{S_j(t_0)}$ is an estimate of $S_j(t_0)$, can be obtained by Kaplan Meier estimate as in 4.4 and $\hat{V}(\hat{S}_j(t_0))$ is an estimate of the variance of the estimated survival function can be obtained by *Greenwood's formula* 4.5.
Then a both sided asymptotic test can be performed accordingly.

From the figure 4.4 we can see that there will hardly be any significance difference between the close bars , so aren't interested in those. Rather we will test those survival probabilities who show a bit difference.

For ODI format		p values
at 0	Rohit vs Shikhar	0.274
	Rohit vs Labuschagne	0.285
at 50	Rohit vs Labuschagne	0.14
	Rohit vs Warner	0.220
at 100	Rohit vs Labuschagne	0.014
	Rohit vs Rahul	0.122

For test format		p values
at 0	Rahul vs Labuschagne	0.183
at 50	Rohit vs Labuschagne	0.023
	Rohit vs Warner	0.102
at 100	Rohit vs Labuschagne	0.15
	Labuschagne vs Rahul	0.172

From the test of significance we can we can see that for ODI Rohit a significantly better probability to survive more than 100 than that of Labuschagne at 5% level as the p value of the test is 0.014. When comparing him to Rahul, who has the second lowest probability of surviving at 100, Rohit does have a higher probability, but the difference between them is not statistically significant even at a 10% significance level.

When comparing the survival probabilities of Rahul with Labuschagne at 0 in the test format, which are the most different, the difference is found to be statistically insignificant at a 10% significance level or lower having p values 0.183. Moving on to 50, the survival probability of Labuschagne is significantly higher than that of Rohit at a 5% significance level having p value 0.023. However, although Warner has a higher probability of surviving 50 runs compared to Rohit, the difference is not statistically significant at a 5% significance level. Finally, at a score of 100, the test of significance for the difference between the best and worst survival probabilities, i.e., between Labuschagne and Rohit, shows that it is statistically insignificant even at a 10% significance level. All the other tests have a p value more than 0.05.

So far, our analysis has focused on the survival probabilities and hazard curves of the batsmen. However, it is important to consider that these probabilities and hazards may vary in different situations for each batsman. In order to gain further insights, we will now investigate how the hazards of the batsmen are influenced by certain factors that are crucial in the tournaments. Specifically, we will study the effects of these factors on the hazards of the batsmen in order to better understand their performance in different scenarios.

5 | Modelling the Instantaneous Failure Rates

Our goal now is to model the hazard rate of the batsmen, taking into account various factors that may influence their likelihood of failure. It is evident that the hazard rates differ among the batsmen and may also vary depending on different factors or situations. To study this, we will utilize a *proportional hazard model*. Here, we will attempt to fit a model that includes the following variables as covariates: 'Format type' (ODI or Test), 'Venue' (Home or Away), and 'Type of innings' (Chasing or Defending). These covariates are expected to have a proportional effect on the hazard rate. Additionally, we will explore an *additive Hazard model*, assuming an additive impact of the covariates on the hazard rate if the proportional model is not good enough.

5.1 | Different types of Regression models

Let, T_i = survival time and C_i = censor time, for $i = 1(1)n$.
with, $y_i = min(T_i, C_i)$ and $\delta_i = I(T_i \leq C_i)$
Also let there be p covariates $\underset{\sim}{x_i} = (x_{1i}, x_{2i}, ...x_{pi})'$
We observe $(y_i, \delta_i, \underset{\sim}{x_i}')$ for $i = 1, 2....., n$

5.1.1 | Proportional Hazard Model

Let, $h(t, \underset{\sim}{x})$: the hazard at time t based on the covariate value $\underset{\sim}{x}$
Then model as, $h(t, \underset{\sim}{x}) = h_0(t).g(\underset{\sim}{x})$
where, $h_0(t) \longleftarrow$ depends on t not on $\underset{\sim}{x}$
and $g(\underset{\sim}{x}) \longleftarrow$ depends on $\underset{\sim}{x}$ not usually on t.
Cox's suggestion: $g(\underset{\sim}{x}) = exp[\underset{\sim}{x}'\underset{\sim}{\beta}]$
Hence the *Cox Proportional model*,

$$h(t, \underset{\sim}{x}) = h_0(t)e^{\underset{\sim}{x}'\underset{\sim}{\beta}} \tag{5.1}$$

if, $\underset{\sim}{x} = 0$ then $h(t, \underset{\sim}{x}) = h_0(t)$, is called the *baseline hazard function.*

Note that 5.1 is a *semi-parametric model*, whose survival function is given by,

$$S(t, \underset{\sim}{x}) = e^{-\int_0^t h_0(u)e^{\underset{\sim}{x}'\underset{\sim}{\beta}}du} = \left[e^{-\int_0^t h_0(u)du}\right]^{e^{\underset{\sim}{x}'\underset{\sim}{\beta}}}$$

$$= \left[S_0(t)\right]^{e^{\underset{\sim}{x}'\underset{\sim}{\beta}}}$$

$S_0(t) \longleftarrow$ baseline survival function.

For estimating the coavriates' effect $\underset{\sim}{\beta}$, we use the *maximum likelihood approach.* Hence the *likelihood function* of the model is given as,

$$l(\underset{\sim}{\beta}, h_0(.)) = \prod_{i \in U} h_0(y_i)e^{\underset{\sim}{x}_i'\underset{\sim}{\beta}} \prod_{i=1}^{n} e^{-H_0(y_i)e^{\underset{\sim}{x}_i'\underset{\sim}{\beta}}}$$

where U is the set of uncensored times and $H_0(t) = \int_0^t h_0(u)du$ so, the *log-likelihood*

$$L(\underset{\sim}{\beta}, h_0(.)) = \sum_{i \in U} log h_0(y_i) + \sum_{i \in U} \underset{\sim}{x}_i'\underset{\sim}{\beta} - \sum_{i=1}^{n} H_0(y_i)e^{\underset{\sim}{x}_i'\underset{\sim}{\beta}} \tag{5.2}$$

From 5.2 it is difficult to estimate $\underset{\sim}{\beta}$, so we maximize a *partial likelihood* that contains the most information about $\underset{\sim}{\beta}$. After few basic algebra we obtain the estimate of β by maximizing the following partial likelihood.

$$l_1((\underset{\sim}{\beta})) = \prod_{i \in U} \frac{e^{\underset{\sim}{x_i}'\underset{\sim}{\beta}}}{\sum_{k \in R_i} e^{\underset{\sim}{x_k}'\underset{\sim}{\beta}}} \tag{5.3}$$

where R_i is the risk set at time y_i.
solving 5.3 in iterative manner, we obtained the estimate as $\underset{\sim}{\widehat{\beta}}$

To test $H_0 : \underset{\sim}{\beta} = \underset{\sim}{\beta}_0$ against any alternative,
can use the *Wald's Statistic*: $(\underset{\sim}{\widehat{\beta}} - \underset{\sim}{\beta}_0)'A(\underset{\sim}{\beta}_0)(\underset{\sim}{\widehat{\beta}} - \underset{\sim}{\beta}_0) \sim \chi^2_p$
where A is the estimated variance matrix for $\underset{\sim}{\widehat{\beta}}$. Also other test can be performed.

To test, significance of a particular parameter, $H_0 : \beta_j = 0$, can use the usual t *test* with test statistic

$$t = \frac{\widehat{\beta_j}}{\widehat{s.e(\beta_j)}} \sim t_{n-p-1}$$

5.1.2 | Check for Proportionality Assumption

The validity of Cox's regression analysis relies heavily on the assumption of proportionality of the hazard rates of individuals with distinct values of a covariate. There are several graphical techniques, one of which is discused here for checking this assumption.
We are interested in checking for proportional hazards for a given covariate x_1 after adjusting for all other relevant covariates in the model, that is, we write the full covariate vector as $\underset{\sim}{x} = (x_1, \underset{\sim}{x_2})'$.where $\underset{\sim}{x_2}$ is the vector of the remaining p-1 covariates in the model. We assume that there is no term in the model for interaction between x_1 and any of the remaining covariates. Assume that the covariate x_1 is categorical and has only K possible values $1, 2, ..., K$. We, then, fit a Cox model stratified on the discrete values of x_1, and let $H_{go}(t)$ be the estimated cumulative baseline hazard rate in the gth stratum. If the proportional hazards model holds, then, the baseline cumulative hazard rates in each of the strata should be a constant multiple of each other. This serves as the basis of the graphical check of the proportional hazards assumption.

To check the proportionality assumption one could plot $log[H_{10}(t)], ..., log[H_{K0}(t)]$ versus t. If the assumption holds, then, these should be approximately parallel and the constant vertical separation between $ln[H_{g0}(t)]$ and $ln[H_{h0}(t)]$ should give a crude estimate of the factor needed to obtain $H_{h0}(t)$ from $H_{g0}(t)$.

5.1.3 | Goodness of Fit: Cox Snell Residuals

The Cox and Snell (1968) residuals can be used to assess the fit of a model.
Let,

$$T \sim S(t) \text{ , } survival\ function$$

Then

$$S(t) \sim U(0, 1)$$

Now, cumulative hazard function $H(t) = -logS(t) \implies S(t) = e^{-H(t)}$
The density of $S(t)$ is given by,

$$f^*(S(t)) = 1 \ , \ 0 \leq S(t) \leq 1$$

Therefore the density of $H(t)$ is,

$$f^{**}(H(t)) = \ 1.jacobian = e^{-H(t)} \ \sim \ Exp(1)$$

so the hazard function for $H(t)$ is

$$\lambda(H(t)) = 1$$

and the cumulative hazard function is given by,

$$\Lambda(H(t)) = \int_0^t 1du = t$$

Hence the plot of $(t, \Lambda(H(t)))$ is a straight line through origin.

Algorithm for Cox-Snell residual plot:

- Model the hazard rate based on suitable covariates and find the estimates of the parameters.
- Obtain the estimated cumulative hazard rates for different time points based on the covariates at that time. The *estimated cumulative hazard rates* are known as the *cox snell residuals*, $r_{cs}(t)$.
- Now use $(r_{cs}(y_i), \delta_i)$ as our data and obtain an estimate of the cumulative hazard at $r_{cs}(y_i)$, in a non-parametric approach.
- Plot the estimate cumulative hazards obtained by non-parametric approach over time.
- For a good fit, the plot will be a straight line through origin with slope 45^0.

5.2 | Covariate Selection

As discussed above we will use three covariates : 'Format type' (ODI or Test), 'Venue' (Home or Away), and 'Type of innings' (Chasing or defending) for modelling the hazard rates of the batsmen as our exploratory analysis suggests. From the data set we will use `'Venue'`, `'chasing'` and `'Type'` variables as my covariates for our model fitting.

5.3 | Hazard Model for Rohit Sharma

The Cox proportional hazards model is a widely used hazard model in survival analysis. In our analysis, we will attempt to fit a Cox proportional hazards model to examine Rohit Sharma's hazard rate, taking into account the selected covariates.

Before proceeding with the model fitting, it is essential to assess the proportionality assumption of the hazard rates based on the covariates. This assumption assumes that the hazard ratios for different covariate values remain constant over time. By examining the proportionality assumption by the method discussed in the section 5.1.2, we can determine the validity of applying the Cox proportional hazards model to our data for Rohit Sharma.

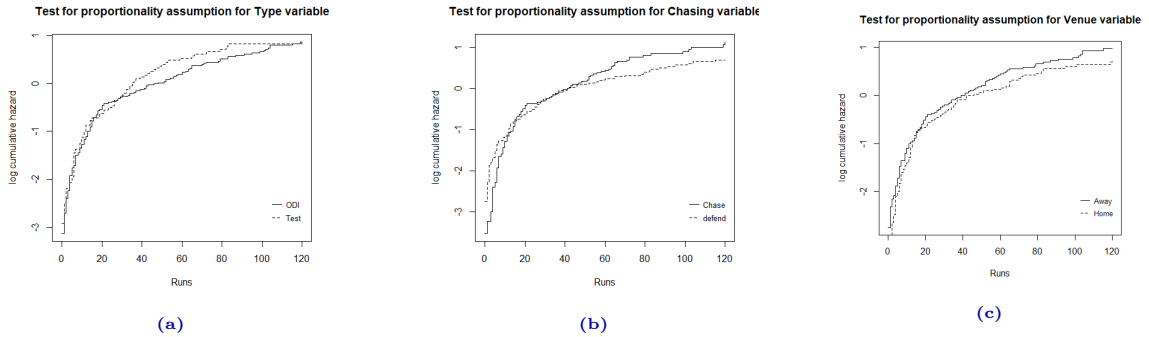


Figure 5.1: Graphical test for proportionality assumption of the covariates — Type, Venue and Chasing on the hazard of Rohit Sharma

Based on the plot, it is evident that the variable "chasing" does not exhibit a proportional effect on the hazard curve of Rohit Sharma. The log cumulative hazard curves show a

complete reversal of their association over time. Initially, the "defend" category has a higher hazard rate, but as the scores increase, the "chasing" category shows a higher hazard rate. This non-proportional effect indicates that the hazard ratio between the "chasing" and "defend" categories is not constant over time. The reversal of the association suggests that the impact of the "chasing" variable on the hazard rate changes as the match progresses or as the scores increase. While the variable "venue" does show a proportional effect. Additionally, the plot of the log cumulative hazards for the "type" variable appears to be close at the beginning but exhibits a good proportional effect later on.

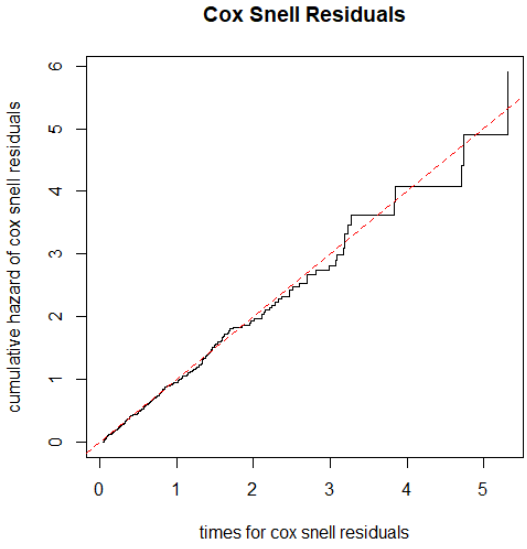
Considering these observations, we will proceed with developing a proportional hazard model for Rohit Sharma based on the "venue" and "type" variables. These variables will be included in the model as covariates to capture their proportional effects on the hazard rate.
Hence fitting a proportional hazard model as per section 5.1.1 we have the following estimates of the parameters

covariate	estimate	exp(estimate)	s.e	p value
TypeTest	0.0945	1.0991	0.1456	0.516
VenueH	-0.2696	0.7637	0.1449	0.062

p value of wald's test = 0.1358

Model Interpretation and Diagnosis

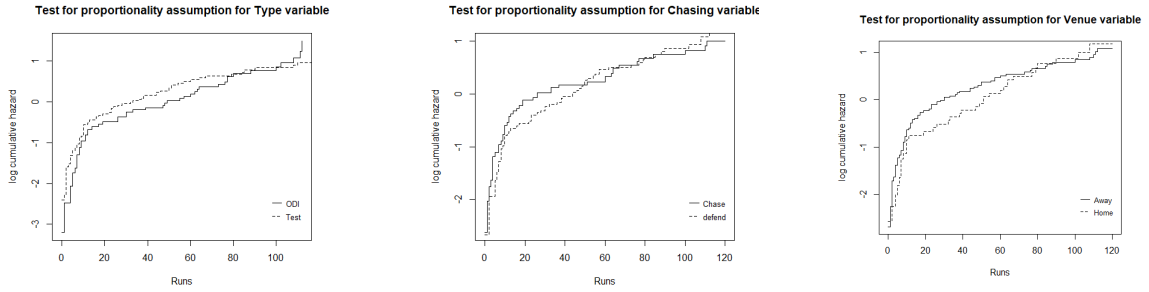
From the model we can say that for Rohit Sharma has a higher hazard in the test format. Coming to the test format from the ODI is hazard treat increases 1.0991 times when the other covariates are fixed. Again he seems to be more stable in the home match because his hazard rate becomes 0.7637 times of the hazard rate of the Away matches irrespective of the format of the game. which implies that he will not be that impactful in WTC as he will be in WC which will be organised in his home turfs.



To check the goodness of fit of the model we use the cox-snell residuals. Although the plot shows that, it is a good model fitting the hazard rates but the effects of the covariates on the hazard of Rohit is too small to be distinguishable and the p values of the estimates say a similar thing that the estimates are too small to have a significant effect on the hazard. Also the p value of the Wald's test says that this model is not at all significantly modelling the hazard rate.

5.4 | Hazard model for KL Rahul

As discussed above, before going to model we need to check the proportionality assumptions of the covariates.



(a) (b) (c)

Figure 5.2: Graphical test for proportionality assumption of the covariates — Type, Venue and Chasing on the hazard of Rahul

Based on the plot, it is evident that the "type" variable and the "venue" variable exhibit a proportional effect on the hazard rates. The log cumulative hazards for both variables are almost parallel, indicating that the proportional hazard assumption holds for these variables. However, it is important to note that in the higher runs, the log cumulative hazards for the "type" and "venue" variables cross each other. While this crossover is not significantly large that's why we assume proportionality, however it suggests some deviation from strict proportionality.
On the other hand, the plot for the "chasing" variable shows that the log cumulative hazards cross each other at multiple points, and the differences between them vary across time. This indicates a violation of the proportional hazard assumption for the "chasing" variable. The crossing of hazard curves suggests that the hazard ratio between the different levels of the "chasing" variable is not constant and changes over time.
Considering these observations, we can conclude that the "type" and "venue" variables have a reliable proportional effect on the hazard rates, while the proportional effect of the "chasing" variable is not reliable due to the violation of the proportionality assumption.

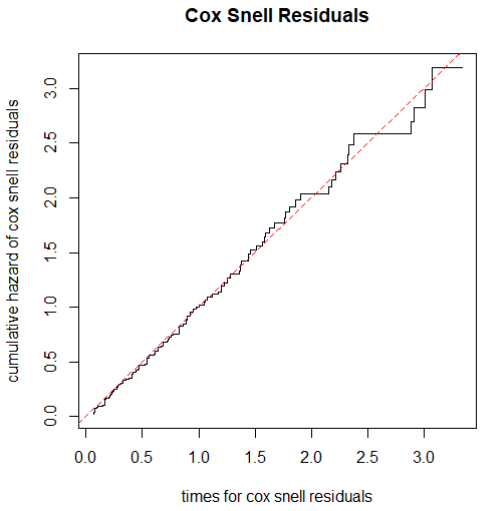
Hence fitting a proportional model we have the estimates for Rahul's data as,

covariate	estimate	exp(estimate)	s.e	p value
TypeTest	0.0822	1.0856	0.1858	0.659
VenueH	-0.1334	0.87507	0.1969	0.498

p value of wald's test = 0.633

Model Interpretation and Diagnosis

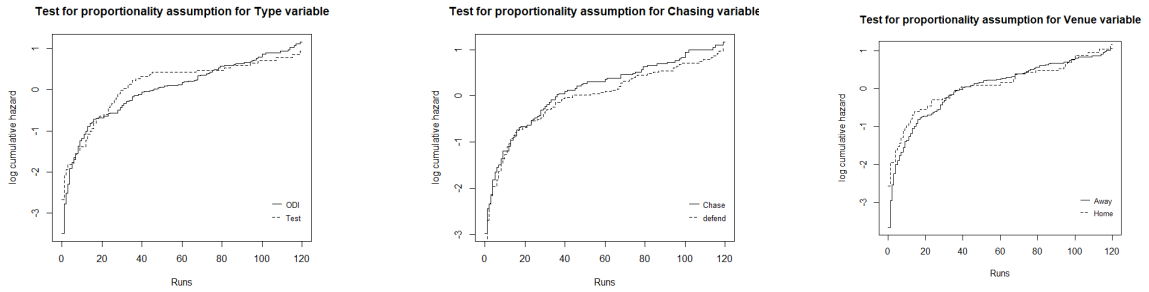
Looking to the estimates for KL Rahul we can assist that in Test cricket he has a higher failure rate which increases 1.085 times of the ODI matches when all the other effects are ignored. Like Rohit when he plays home matches he is less likely to be out as the estimate says in home matches his hazard rate drop by 0.87507 times than the away matches irrespective of the format. So his impact can be better experienced in world cup matches.



However, in the case of KL Rahul, the hazard estimates are very close to 0, suggesting that the differences in hazard rates based on the covariates may not be significantly large. The p-values associated with the corresponding covariates also support this observation, as they fail to reject the null hypothesis of statistical insignificance at the 5% level. Additionally, even a good fit of Cox-Snell residuals is not meaningful in this case, as the Wald test indicates that the model is statistically insignificant even at the 5% or 10% level.

5.5 | Hazard Model for Shikhar Dhawan

Begin with the proportionality check like the previous cases.



(a) (b) (c)

Figure 5.3: Graphical test for proportionality assumption of the covariates — Type, Venue and Chasing on the hazard of Dhawan

Studying the graph, we observe that for Shikhar Dhawan, the "Type" variable does not exhibit a proportional effect on the hazard curve. Between the scores of 20 to 80, there is a sudden increase in the difference of the log cumulative hazard curves, indicating a violation of the proportional hazard assumption. However, at the beginning and above 80, the difference is relatively small.
On the other hand, for the other two variables, we can assume a proportional effect on the hazard curve. Although there may be some crossover among the log cumulative hazard curves,as the differences are relatively low.

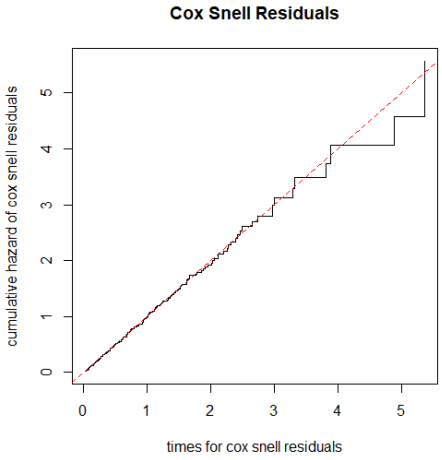
Now we will fit a proportional model using 'venue' and 'chasing' on Shikhar's data.

covariate	estimate	exp(estimate)	s.e	p value
VenueH	0.065	1.067	0.15	0.668
chasing d	-0.0488	0.9523	0.1387	0.725

p value of wald's test = 0.8615

Model Interpretation and Diagnosis

Dhawan doesn't have any proportional effect of format of the game to his hazard. So forgetting that for we observed that he is less frail in the away matches when the innings type is not considered. His hazard rate in home matches becomes 1.06 times of the hazard rate of the away matches. He also has a effect of the type of innings in his hazard. While batting first he is likely to survive more on the crease than while chasing.

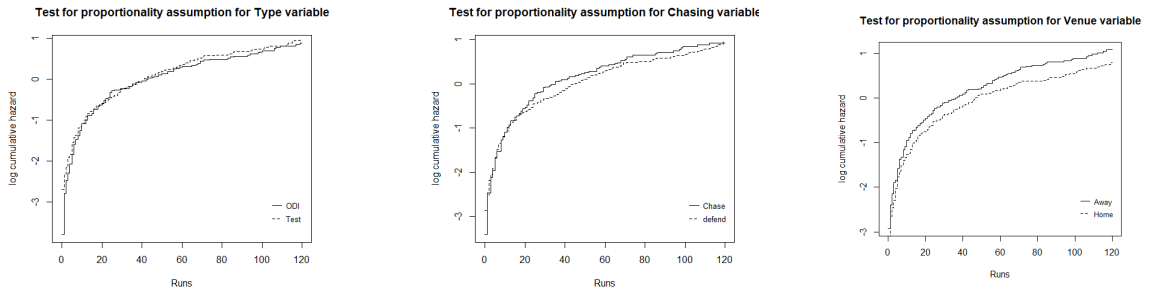


similar inference can be drawn for this case as well. The estimated coefficients for the variables are statistically insignificant at the 5% level, indicating that there is no significant relationship between the covariates and the hazard rate. The p-values of the Wald's test further support this observation, suggesting that the model is not statistically significant at either the 5% or 10% level. Therefore, the good fit of the Cox-Snell residuals may not be meaningful in this context.

A

5.6 | Hazard Model for David Warner

An interesting observation can be made regarding David Warner's hazard analysis. Unlike the previous three batsmen, Warner's hazard demonstrates a proportional effect of all the covariates.



(a) (b) (c)

Figure 5.4: Graphical test for proportionality assumption of the covariates — Type, Venue and Chasing on the hazard of Warner

Starting with the 'venue' variable (Figure c), we can observe that the two hazard curves are parallel, indicating the presence of a proportional effect of the venue variable on the hazard.

Moving on to the 'chasing' variable, the two hazard curves are very close at the beginning but start to differ after 20 runs, increasing in a parallel manner. This suggests a proportional effect of the chasing variable on the hazard.

Lastly, for the 'Type' variable, the two hazard curves are almost superimposed upon each other, indicating that the difference in hazard between the categories of this variable will be small. This further supports the assumption of a proportional effect.

Overall, these observations suggest that the hazard rates of David Warner are influenced by the different covariates in a proportional manner, as evidenced by the parallel nature of the hazard curves.

Hence on fitting cox proportional model based on all the covariates we have the following estimate for David Warner.

covariate	estimate	exp(estimate)	s.e	p value
TypeTest	0.0582	1.0598	0.1154	0.614
VenueH	-0.31023	0.733	0.1176	0.008
chasing d	-0.0451	0.955	0.1182	0.702

p value of wald's test = 0.0518

Model Interpretation and Diagnosis

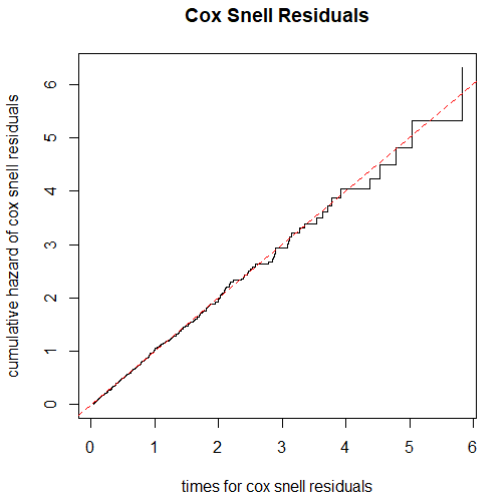
Talking about David Warner when it is the test format his hazard rate increases 1.0598 times of the hazard rate of ODI format when all the other parameters of the match is fixed. While playing in home situation he his failure rate becomes 0.733 times of failure rate in away situations. Even his failure rate is significantly lower in home situation than in away at 5% and 1% level. While giving a target to the opponent he is least frail in that innings then chasing a score. his hazard rate while batting first is 0.95 times of the hazard rate while chasing but this difference is not statistically significant at 5% level. The wald test says that the model is acceptable to model his hazard at 10% level with a p value of 0.0518 and model AIC 2984.303.

On looking for a better model, when we drop the *chasing* variable, having a higher p value, the p value of the wald test becomes 0.022 which indicates that the model is statistically significant in modeling the hazard rate of David Warner at 5% level and model AIC: 2982.468 .

covariate	estimate	exp(estimate)	s.e	p value
TypeTest	0.0604	1.0622	0.1152	0.6
VenueH	-0..317	0.722814	0.11161	0.0063

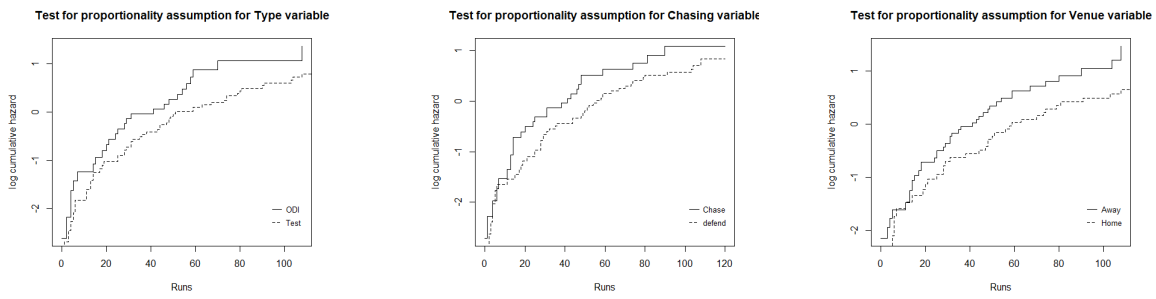
p value of wald's test = 0.0022

The second model indicates that in home turfs his hazard rate is 0.7228 times than that of in away. Even the coefficient being statistically significant at 5% level we can notice a significant difference between his home and away performance. In test format his hazard his 1.06 times of the hazard rate in ODIs. If we try to give a prediction based on the model we can say that he will not that much effective in the tournaments as he used to be as the tournaments are in away turf for him and comparing the format, a slightly better performance can be seen in the ODI world cup.



To check the goodness of fit of the model we plotted the cox snell residuals of the second model along with time and it is almost a 45^0 straight line from the origin, suggesting a well fitting. Also the model AIC decreases.

5.7 | Hazard Model for Labuschagne



(a) (b) (c)

Figure 5.5: Graphical test for proportionality assumption of the covariates — Type, Venue and Chasing on the hazard of Labuschagne

The hazard analysis for Labuschagne reveals that the log cumulative hazard curves for the different categories of the covariate variables, namely 'type', 'chasing', and 'venue', are parallel. This indicates that Labuschagne's hazard is proportional to these covariates. Given the parallel nature of the log cumulative hazard curves, it is reasonable to fit a proportional hazard model for Labuschagne using these covariates. The assumption of proportionality holds for Labuschagne, suggesting that the hazard rates vary in a proportional manner based on the covariate categories.

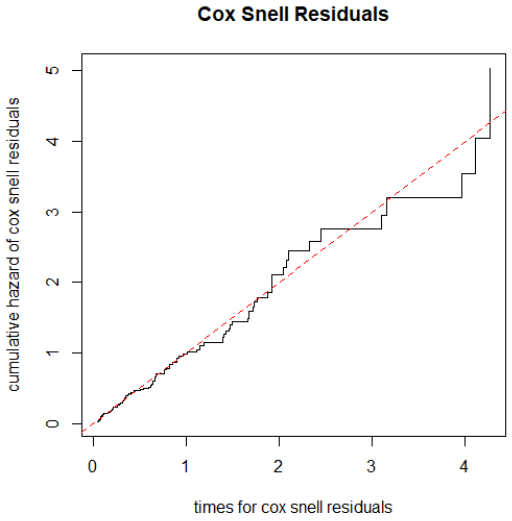
Hence the estimates of the proportional model is,

covariate	estimate	exp(estimate)	s.e	p value
TypeTest	-0.4477	0.6391	0.2404	0.06
VenueH	-0.4355	0.6469	0.2533	0.085
chasing d	-0.4204	0.6568	0.2342	0.072

p value of wald's test = 0.0060

Model Interpretation and Diagnosis

From the estimated coefficient of the cox proportional model we can say that Labuschagne's hazard in test matches is 0.6391 time than the ODIs. Getting the home situation he is more successful than away situations as his hazard in home is 0.6469 time of the hazard in away. While batting first his hazard is 0.6568 times of the hazard while chasing. i.e. he keep more potential in score in the 1st innings.



From the graphs in Figure 5.5, we can observe that the two log cumulative hazard curves maintain a relatively constant difference throughout the entire time period. This pattern is more pronounced for Labuschagne compared to the other batsmen. However, despite the estimates being more effective in distinguishing the hazard rates for Labuschagne, they are not statistically significant at the 5% level. One possible reason for this could be the relatively smaller number of matches played by Labuschagne compared to the other batsmen in the analysis. Nevertheless, the coefficients show significance at the 10% level.

Additionally, the p-value of the Wald test is 0.0060, indicating that the model fitting is statistically significant at the 5% level for modeling Labuschagne's hazard.

Impact in the tournaments

Speaking about the tournaments, Labuschagne is expected to have a significantly better performance in the WTC final against India, as he demonstrates a stronger ability to survive in a Test innings compared to an ODI. His analysis of hazard rates and survival probabilities suggests that he has a higher chance of performing well in the Test format, indicating that his skills are better suited for longer matches. Therefore, Labuschagne's performance in the WTC final is expected to be more impressive than in an ODI match.

Taking about the ODI world cup, Rohit, the man will be stand out then than others, as he will get the benefits of home matches where his hazard rate is comparatively lower and the format suits him for a better survival.

Indeed, the proportional hazard model has provided valuable insights into the effects of different factors on the hazard rates of the batsmen so we are not going towards *additive hazard model*. As we are analyzing the performance of the batsmen for the upcoming tournaments and aiming to understand their hazard rates, it would be beneficial to develop a model to estimate their expected scores in these tournaments.

By incorporating relevant covariates such as venue, type of match, and chasing/defending, along with the individual player's characteristics, we can construct a predictive model for estimating the expected scores of the batsmen. This model can take into account the hazard rates derived from survival analysis and provide valuable information

on the potential run-scoring abilities of the batsmen in the tournaments.

With such a model, we can gain a deeper understanding of the expected performance of the batsmen, allowing us to make more informed decisions and predictions for the upcoming tournaments.

6 | Modelling the Runs of the batsmen

Certainly, developing a model to predict the runs scored by batsmen based on covariates such as format type, venue, and type of innings can provide a better insight into their performance.
By considering these factors, we can capture the variations in runs scored under different match situations. By analyzing historical data and incorporating the relevant covariates, we can develop a model to predict the expected runs scored by batsmen in 'different formats' (ODI or Test), 'venues' (Home or Away), and 'types of innings' (Chasing or Defending). The model can then provide valuable insights into the performance of the batsmen and help in making informed decisions about team selection, strategy, and match expectations.

It is important to note that the accuracy and reliability of the model will depend on the quality and availability of the data, as well as the suitability of the chosen regression technique.

6.1 | Accelerated Failure time model

The *accelerated failure time(AFT) model* is a parametric survival model that the model the log of the event time by an linear regression manner.

With our observe data: $(y_i, \delta_i, \underset{\sim}{x_i}')$ for $i = 1, 2....., n$, we model the time of the event as,

$$log(y_i) = \underset{\sim}{x_i}' \underset{\sim}{\beta} + \sigma\varepsilon_i \quad i = 1(1)n \tag{6.1}$$

with

$$\varepsilon_i \sim f(\varepsilon_i) \;\; with \;\; E(\varepsilon_i) = 0 \;\&\; Var(\varepsilon_i) = 1$$

For different choices of $f(.)$, we have different Aft models.
Usually based on the distributional assumption of T, time to the event we have choices of $f(.)$. Since T is non negative, distribution choice of T is not normal like usual linear model.

6.2 | Fitting of AFT model

In our analysis time of the event is the 'runs a batsman score before getting out', and the covariates are : 'Format type' (ODI or Test), 'Venue' (Home or Away), and 'Type of innings' (Chasing or defending)

In the context of survival analysis the commonly used distribution of the event time are — Log-logistic, Weibull, Log-normal, Gompartz.
To choose a suitable Aft model, we fitted the model for these four distribution for each

of the batsmen and choose one which has a lesser AIC.

Below in the table the AIC of the alternative models are shown for each of the batsmen.

Distribution	Rohit	Rahul	Dhawan	Warner	Labuschagne
llogis	2047.52028	1089.02747	2056.20305	3017.01660	864.37116
Weibull	1979.64170	1052.73885	1996.45956	2915.20109	827.59819
lnorm	2134.80663	1126.60799	2138.22377	3141.89591	903.08330
gompatz	2004.86800	1096.99694	2020.16352	2957.94440	830.36068

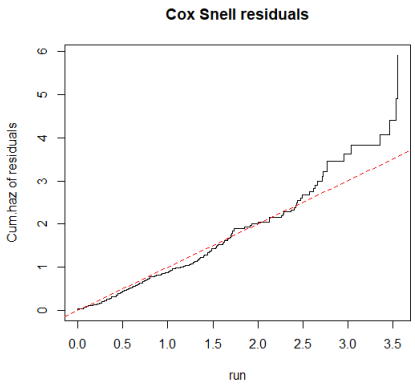
Hence for each of the batsman the weibull Accelerated failure model has the lower AIC. So, we will fit weibull AFT model for the batsman to have a look the effects of the covariates in their scoring runs.

The weibull AFT model also has another property — fitting this model we actually reduce to the cox proportional hazard model, if we study the characteristic of the hazard by the model. This is another aspect of AFT model but we will not look into that in our analysis. However the lower AIC of the weibull indicates that modelling the hazard thorough the proportional model was a good thinking.

Weibull AFT model for Rohit

Fitting the accelerated model for Rohit Sharma we have the output:

coefficient	**Estimates**	**pvalue**
Intercept	3.4349	<2e-16
TypeTest	-0.1502	0.43
VenueH	0.2629	0.15
chasing d	0.2223	0.22
log(scale)	0.2717	

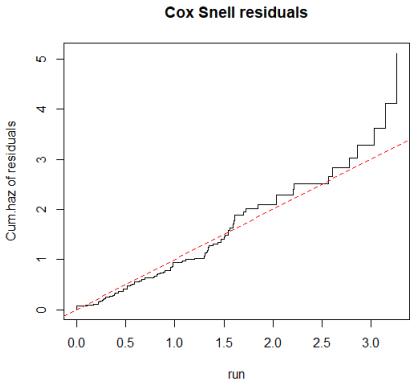


When the format changes from ODI to test the log runs of Rohit Sharma lower down by 0.1502 units When the other factors are fixed. Playing in home situation and batting in the first innings his log run increases by 0.2629 and 0.2223 units respectively considering one factor at a time and keeping other fixed.
Although the increments or decrements in log run is not statistically significant at 5% level. But an association of the log run on the covariates can be studied from the model. Cox snell residuals plot for the model is given which show that the model fits the data well.

Weibull AFT model for Rahul

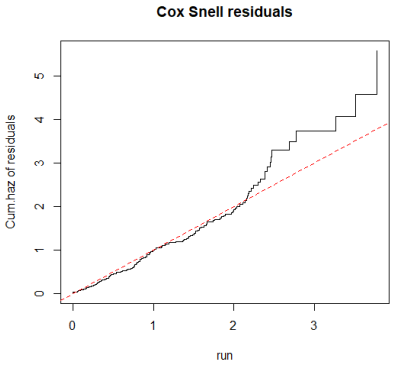
A similar result can be noticed for Rahul also. In the test format the log run of Rahul decrease by -0.4114 units from ODI. playing in the home condition his log run increases by 0.35034 units. The effect of innings type to the log run is almost negligible, as log run gets better 0.0078 when the team is batting first.

coefficient	Estimates	pvalue
Intercept	3.5645	<2e-16
TypeTest	-0.4114	0.19
VenueH	0.35034	0.28
chasing d	0.0078	0.98
log(scale)	0.4793	



Weibull AFT model for Dhawan

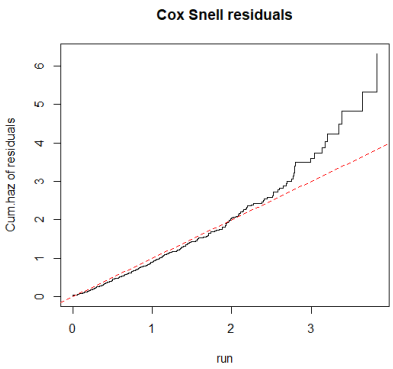
coefficient	Estimates	pvalue
Intercept	3.7001	<2e-16
TypeTest	-0.1323	0.52
VenueH	-0.0711	0.72
chasing d	0.0417	0.82
log(scale)	0.265	



In contrary to Rohit and Rahul the log run of Shikhar increases in away condition by 0.0711 units indicates his better scoring potential outside India. The cox snell plot gives us a challenge here as at the higher value it is not that good fit.

Weibull AFT model for Warner

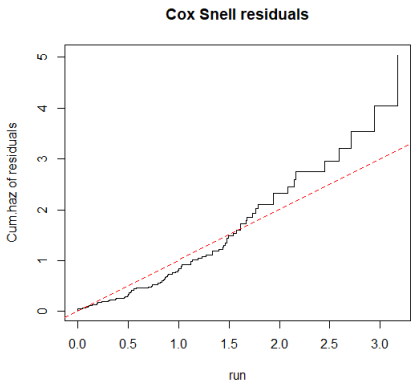
coefficient	Estimates	pvalue
Intercept	3.4273	<2e-16
TypeTest	-0.0786	0.608
VenueH	0.3203	0.037
chasing d	0.072	0.647
log(scale)	0.2885	



A very similar result like Rohit and Rahul can be stated for David Warner as he has more potential score better in ODI then test with a change of 0.0786 units log runs. unlike other he is significantly better at 5% level in scoring in Australia with a change of 0.3203 in log run than away conditions.

Weibull AFT model for Labuschagne

coefficient	Estimates	pvalue
Intercept	3.0226	<2e-16
TypeTest	0.3852	0.197
VenueH	0.4035	0.15
chasing d	0.3436	0.219
log(scale)	0.1985	



Labuschagne show a a opposite association in scoring runs in test than his competitor. His log runs in test format increases by 0.3852 units than in ODI. Though the estimate is not statistically significant, he keep potential to score better in longer format of cricket. While home condition his log runs increase 0.403 units than away condition. Also while batting first, his log runs increase by 0.3436 units.

Impact in tournaments:

Based on this AFT models if we look for the best performer ICC tournaments among these 5 batsmen, we can say that Rohit Sharma keeps potential to score a better run in the ODI World Cup matches rather than the world test championship final, as the conditions of World Cup suits him better then world test championship final to score a better run. Whereas for Labuschagne the longer format of cricket suits him better than ODI. As the test championship final will be played oppose to the home conditions of the batsman nobody would get benefited unless Dhawan, but unfortunately he is not is the squad of the test championship final. As a result, Labuschagne is considered the most likely to score more runs among the given batsmen in the Test Championship final. In the ODI format, Rohit Sharma and KL Rahul are predicted to have better scores due to the advantage of playing on home turfs, while David Warner and other batsmen may not benefit as much from the conditions.

It's important to note that these predictions are based on the assumptions and opinions provided and may not necessarily align with the actual performance of the players in the respective tournaments. Various factors such as form, fitness, team dynamics, and match conditions can significantly impact the performance of individual players in cricket.

7 | Competing Risk model

Up until now, our analysis has focused on the runs scored by a batsman before being out in any form. However, in cricket, there are multiple ways in which a batsman can be dismissed, such as being caught, bowled, given out LBW, or run out. Now, we are specifically interested in the event of being out by either being caught, bowled, LBW, or run out, but we only observe the runs scored before the first occurrence of such an event in a particular innings. In other words, we are now considering a "*cause-specific event*" where the cause of interest is being out by one of the mentioned methods.

While a marginal analysis can be performed by considering a specific reason for dismissal as the event of interest and treating other reasons as censored, this approach relies on a strong assumption that the censoring is independent of the event. However, in the presence of competing risks, this assumption of independence becomes questionable. Therefore, a more appropriate analysis would involve considering the competing risks together. In the presence of competing risks, interpreting survival analyses becomes inherently ambiguous due to the uncertainty surrounding the degree of dependence among the competing outcomes. This means that the results and conclusions drawn from survival analysis in the presence of competing risks should be interpreted with caution, acknowledging the potential for ambiguity due to the uncertainty surrounding the dependence among the different types of events.

7.1 | Cause-Specific Cumulative Incidence Functions

To develop a formal model to accommodate competing risks, let us suppose that there are K distinct causes of event.
The distinguishing feature of this competing causes framework is that each subject can experience at most one of the K causes of death; the times that the subject would have experienced the remaining causes is thus unknown.

With competing risks, it is helpful to define, for each cause of interest, a function known as the *cumulative incidence function*. This is the cumulative probability that an individual dies from that particular cause by time t, and is given by

$$F_j(t) = Pr(T \leq t, C = j) = \int_0^t h_j(u)S(u)du \tag{7.1}$$

Where $S(t)$ is the survival probability at t, and $h_j(t)$ cause specific hazard rate due to j th reason.

Suppose now that we have D distinct ordered failure times $t_1, t_2, ..., t_D$.The cause-specific hazard for the jth hazard may be written as

$$\hat{h_j(t_i)} = \frac{d_{ij}}{n_i}$$

where, $n_i = \#$ at risk at t_i
$d_{ij} = \#$ event j at time t_i
$r_{ij} = \#$ event due to other causes at time t_i

Then the estimated *cumulative incidence function* for the *jth* event is given by,

$$CI_j(t) = \sum_{t_i \leq t} \left[\prod_{k=1}^{i-1} \left(1 - \frac{d_{kj} + r_{kj}}{n_{kj}}\right) \right] \frac{d_{ij}}{n_{ij}} \tag{7.2}$$

7.2 | Estimated CIs for the Batsmen

In this analysis, our focus is to examine the different modes of dismissal for batsmen in various formats of cricket. By gaining insights into the causes of their dismissals, we can better understand why a batsman may have been vulnerable or struggled in a particular situation or innings.

For our analysis, we specifically consider three modes of dismissal: "caught," "bowled," and "lbw." We treat all other modes of dismissal, such as "run out," as censored innings due to the limited number of observations.
To explore the causes of dismissals, we estimate the cumulative incidence functions for the three modes using the equation referenced as 7.2. The resulting functions are depicted in Figure 7.1 for the two cricket formats under discussion.

Upon analyzing Rohit Sharma's batting performance, it becomes evident that he is more prone to being caught out in both formats of cricket. This observation is supported by the purple curves, which indicate that he is equally susceptible to getting caught out early in his innings. However, there is a slight decrease in the probability of being caught out after spending 20 runs in the limited-overs format, which is not as apparent in his Test innings. In terms of being dismissed leg before wicket (lbw), Rohit Sharma has a lower probability of being given lbw in ODI cricket compared to Test cricket. This discrepancy is highlighted by the contrasting trends in the respective orange curves. It suggests that he is more likely to be dismissed lbw in the longer format of the game. As for being bowled, the red curves do not exhibit any significant differences,though in test his has a greater probability of being bold in the early runs. So getting a good start Rohit uses that to make better score.

Talking about Rahul, his probability of being caught out is very high test format than the ODI as the purple curve for test format has a very high slope at the beginning of the innings and slope decreases a bit only after getting 30s whereas if we look at the purple curve for ODI format it has smaller slope than the test's indicates that a lesser probability of being caught out in ODIs. Talking about bowled, he has a higher probability of being bowled in test matches with a very high slope at the beginning indicates instability at the starting of his innings in that format. He also has higher slope on being lbw at the beginning of his test innings.

Coming to Dhawan's analysis we can see that the purple line in test format has a higher slope in the beginning than ODI, indicates his high chance of being out early in test matches. Talking about other two reason of being out he is more likely to be lbw than bowled in test cricket whereas in ODI it is equal.

Talking about Warner almost have same probability of being out in test and ODI by caught and bowled although, his probability of being lbw in ODI cricket is much lesser.

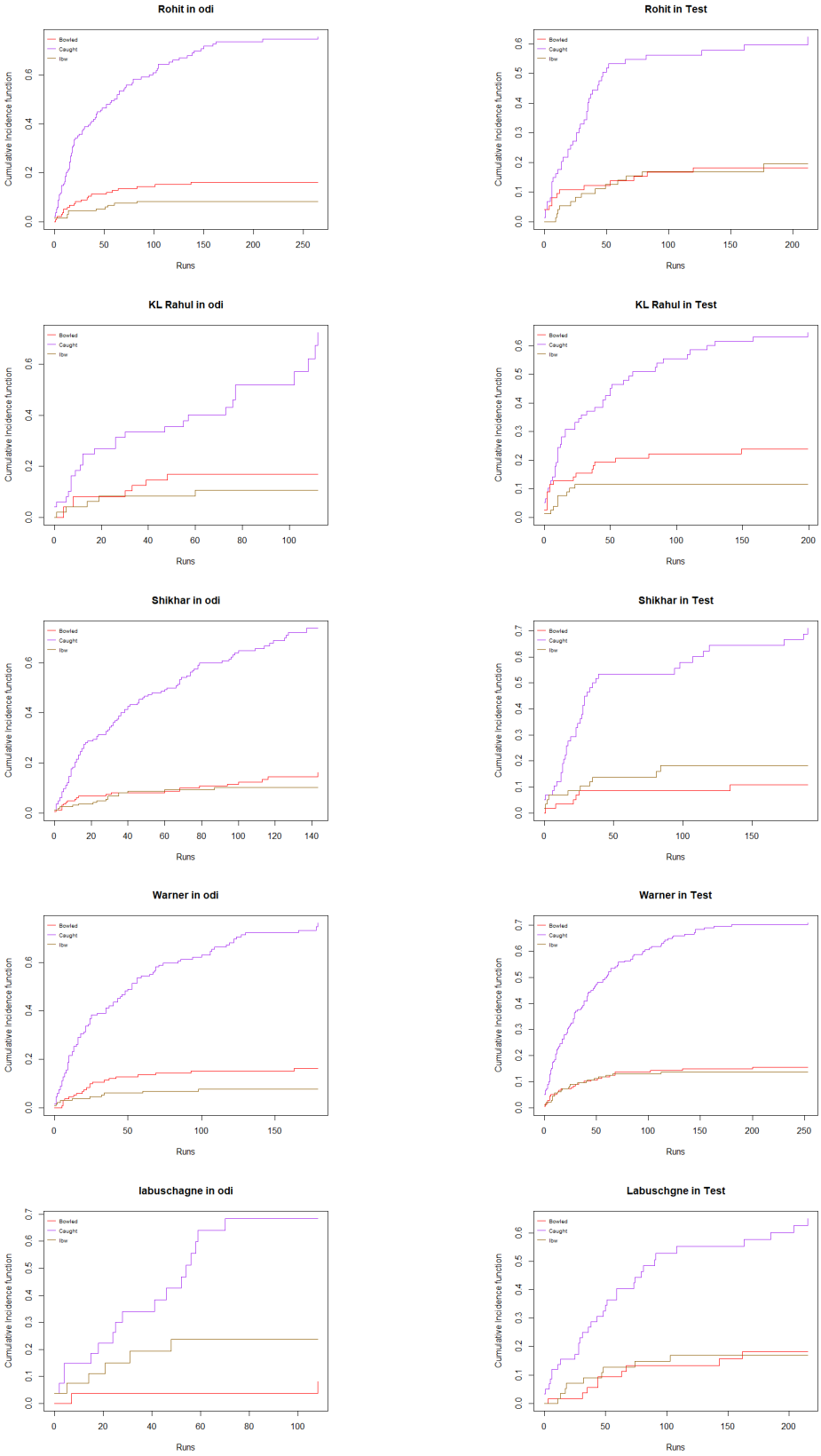


Figure 7.1: Estimated Cumulative incidence functions of the batsman for two format of the game

A reason may be that in ODI he plays like an explosive batsman so probability of being lbw is lower and caught out is higher and almost stays same.

Talking about Labuschagne he has higher probability of being lbw in ODIs than in test—suggesting his awareness of the longer format of the game. Unlike other an interesting observation can be made for Labuschagne is that in test format the slope of the curves are much smaller than the other batsmen indicating his stability at the early stage of his innings that makes him so successful in that format.

8 | Conclusion of the Study

Now its time to draw a overall conclusion of our analysis. We start this analysis with an objective to find impactful players for the two up coming ICC tournaments.

Our exploratory analysis provides us with an initial understanding of batsmen who have displayed a higher average and good recent form. However, it is important to note that the calculated average might be underestimated due to the presence of censor data. In light of this, Labuschagne emerges as a player to watch in the World Test Championship final. His impressive batting average and extended time spent at the crease make him a standout performer. On the other hand, Rohit Sharma has shown exceptional form in recent years, making him the standout player for the upcoming World Cup.

To gain a more comprehensive understanding, we conducted a survival analysis on the batsmen's performance data. By examining the hazard curves and survival probabilities at critical stages of an innings, we determined that Labuschagne is expected to be the top performer in the test format. On the other hand, Rohit Sharma and David Warner are projected to excel in the ODI format.

Since different factor of a match may effect the hazard of the batsman we try to develop a cox proportional model. Keeping the situations of the test championship final and world cup in mind we find that Labuschagne is likely to survive better in the test championship final. Shikhar who a lower hazard in away situations, played his last test match in 2018 and is not in the squad. Getting the home situation in favour Rahul will suppress Warner in World Cup 2023 after Rohit.

Getting a idea about effects of different situation in scoring, we have that Labuschagne will have a better score in test championship final whereas Rohit will get good scores in ODI world cup than others and strictly followed by Rahul, having a good recent form as well as home turfs.

In terms of different modes of dismissal, Labuschagne demonstrates greater stability in the long format of the game compared to the other players. On the other hand, Rahul exhibits strong performance in ODI matches. However, when it comes to getting off to a solid start, Rohit and Warner outshine Rahul in the ODI format.

So at the end if we have to pick the players to look in these ICC tournaments then it will be Labuschagne in WTC final and Rohit in World cup. However they will getting a

close competition there by the others as well.

Limitation

The study has a few limitations that need to be acknowledged.

- Firstly, the “average score" of the batsmen may not accurately represent their true capabilities. While we have remove the censoring effects, the calculated average is likely to be a *underestimate* of their actual batting average.
- Secondly, this analysis does not consider the current fitness status and form of the batsmen. Their selection for the upcoming tournaments may be influenced by fitness issues, which could raise questions about their performance. KL Rahul, who would be an obvious choice for India's side, will be absent from the WTC final due to injury.

9 | References

[1] Per Kragh Andersen and Richard D Gill. Cox's regression model for counting processes: a large sample study. *The annals of statistics*, pages 1100–1120, 1982.

[2] John D Kalbfleisch and Ross L Prentice. *The statistical analysis of failure time data*. John Wiley & Sons, 2011.

[3] John P Klein and Melvin L Moeschberger. *Survival analysis: techniques for censored and truncated data*. Springer, 2003.

[4] David G Kleinbaum and Mitchel Klein. *Survival analysis a self-learning text*. Springer, 1996.

[5] RG Miller. *Survival Analysis*.

[6] Dirk F. Moore. *Applied Survival Analysis using R*. Springer.

A | Appendix: R codes

```
# Libraries
library(dplyr) ; library(mstate)
library(survival) ; library(ggfortify)
library(ggplot2) ; library(ggthemes)
library(flexsurv) ; library(ggmap)
library(ggthemes) ; library(mice)

#========== my functions
# For smooth hazard curve
smooth_haz<- function(model, colour){
  km_surv<- model$surv    # values of the survival functions
  a<-(diff((-log(km_surv)),1) )
  km_haz<-c( (-log(km_surv))[1] ,a )
  mx=which.max(model$time)-1
  x= model$time
  y= km_haz
  plot(x,y,col=NA,xlab = "Runs", ylab = "Hazard rate",
       xlim=c(0,120), ylim = c(0,0.7)
       )
  lines(smooth.spline(x[-length(x)],y[-length(y)],spar = 0.5))
}
#function to make 120+ score censored
make.cens<- function(my_player){
  for(i in 1:nrow(my_player)){
    if(my_player$Runs[i]>120){
      my_player$r[i]= 120
      my_player$dx[i]=0
      }
    else{
      my_player$r[i]= my_player$Runs[i]
      my_player$dx[i]=1
    }
  }
  return(my_player)
}
#========== Web scrapping
library(rvest)
library(xml2)
library(tidyverse)
url_player="url_of_the_data"
page <- read_html(url_player)
col_table<-page %>%  html_nodes("table.engineTable") %>%  html_table() %>% .[[4]]
View(col_table)
write.csv(col_table,"player_name.csv")

#========== Importing the data and preprocessing
df_player<- read.csvl("player_name.csv")
df_player %>%
  group_by(Type) %>%
  summarise(length(Type))
# removing the t20 innings
df_player<- subset(df_player, df_player$Type != "T20I")
# creating delta variable
df_player$delta<- ifelse(df_player$Dismissal=="not out",0,1)
summary(df_player)          # check NA values
```

```
df_player<- na.omit(df_player)  # remove the NA values
dim(df_player)              # check the dimension of the data
# creating chasing
df_player$chasing<- ifelse(df_player$Inns %in% c("2","4"),"c","d")
                            # c: chasing , d: Defending
# lets have a look to the df
head(df_player)
# Data For survival analysis
my_player<- data.frame("Runs"= df_player$Runs, "delta"=df_player$delta,
                   "Type"=as.factor(df_player$Type),
                   "Opposition"= as.factor(df_player$Opposition),
                   "Vanue"=as.factor(df_player$Vanue),
                   "Chasing"=as.factor(df_player$chasing)
                  )
dim(my_player)
my_player<- make.cens(my_player)
dim(my_player)
summary(my_player)
#==================== Survival Analysis
# Survival Curve fit for kl
player.odi<- survfit(Surv(r, dx)~1, data=subset(my_player,Type=="ODI"))
player.test<- survfit(Surv(Runs, delta)~1, data=subset(my_player,Type=="Test"))
#survival curve
ggsurvplot(survfit(Surv(r, dx)~(Type), data=my_player),
           data=my_player, pval=TRUE, conf.int = TRUE,size=1,
           palette=c("#8F43EE","#0E8388"), conf.int.alpha=0.2
)+
  labs(title = "player name: Survival Curve for different format")

survdiff(Surv(r, dx)~(Type), data=my_player) # MH test
# smoothed hazard function
smooth_haz(player.odi)
title(main="Player Name")
smooth_haz(player.test)
title(main="Player Name")
#======== Survival Prob. at specific time
# for odi at t
a1<-kl.odi$surv[max(which(kl.odi$time==t))]
a1s<-summary(kl.odi)$std.err[max(which(kl.odi$time==t))]
a2<-rs.odi$surv[max(which(rs.odi$time==t))]
a2s<-summary(rs.odi)$std.err[max(which(rs.odi$time==t))]
a3<-sd.odi$surv[max(which(sd.odi$time==t))]
a3s<-summary(sd.odi)$std.err[max(which(sd.odi$time==t))]
a4<-dw.odi$surv[max(which(dw.odi$time==t))]
a4s<-summary(dw.odi)$std.err[max(which(dw.odi$time==t))]
a5<-ml.odi$surv[max(which(ml.odi$time==t))]
a5s<-summary(ml.odi)$std.err[max(which(ml.odi$time==t))]
m<- cbind(rbind(a1,a2,a3,a4,a5),rbind(a1s,a2s,a3s,a4s,a5s))
#for odi we find the survival probabilities at t=0,50,100
#for the players and store them in a matrix m1
rownames(m1)<- c("KL","RS","SD","DW","ML")
colnames(m1)<- c("at 0","se0","at 50","se50","at 100","se100")
barplot(t(m1), beside = T,border = NA, col=c("#4E944F","#83BD75","#B4E197"))
# Then we test for their equality by the formula discussed in the section 4.4.1
# The same is done for test innings

#======= Proportional Hazard model
# check for proportionality of the hazard for KL
```

```
# For Venue
player.ph.diag<- coxph(Surv(r,dx)~strata(Vanue),data=my_player, method = "breslow")
df.base.haz<-basehaz(player.ph.diag)
diag.a<- df.base.haz %>%  filter(strata=="A")
diag.h<- df.base.haz %>%  filter(strata=="H")
plot( diag.a$time, log(diag.a$hazard), type="s", xlab="Runs", ylab="log cumulative hazard",
main="Test for proportionality assumption for Venue variable" )
lines( diag.h$time, log(diag.h$hazard), type="s", lty=2 )
legend("bottomright", legend = c("Away","Home"),lty=c(1,2),bty="n", inset = 0.05, cex=0.8 )
# same is done for other two covariates
cox.zph(coxph(Surv(r, dx)~Type+Chasing+Vanue,data=my_player, method="breslow"))
# check propotionality assumption by Schoenfeld residuals
# Now the coxph model
player.ph<- coxph(Surv(r, dx)~Type+Vanue,data=my_player, method="breslow")
player.ph
# cox snell residuals
my_player$mart<- residuals(player.ph, type = "martingale")
my_player$cox_snell<- -(my_player$mart - my_player$dx)
player.coxsnell<- survfit(Surv(cox_snell,dx)~1,data=my_player)
player.coxsnell$cumhaz
plot(player.coxsnell$time,player.coxsnell$cumhaz,  pch=19,type="s",
     main="Cox Snell Residuals", xlab = "times for cox snell residuals",
     ylab="cumulative hazard of cox snell residuals")
abline(0,1, lty=2, col="red")

#======== AFT models
aft.aic<- function(df){
bbb<-sum(df$Runs==0)
df$Runs[which(df$Runs==0)]<- rep(0.001,bbb)
mll<- flexsurvreg(Surv(Runs, delta)~ Type+ Vanue + Chasing, data= df, dist = "llogis")
mw<- flexsurvreg(Surv(Runs, delta)~ Type+ Vanue + Chasing, data= df, dist = "weibull")
mln<- flexsurvreg(Surv(Runs, delta)~ Type+ Vanue + Chasing, data= df, dist = "lnorm")
mg<- flexsurvreg(Surv(Runs, delta)~ Type+ Vanue + Chasing, data= df, dist = "gompertz")
a1<- c("llogis","Weibull","lnorm","gompatz")
a2<- c(AIC(mll),AIC(mw),AIC(mln),AIC(mg))
names(a2)<- a1
return(a2)
}
www<-t(rbind(
  aft.aic(my_rs),
  aft.aic(my_kl),
  aft.aic(my_sd),
aft.aic(my_dw),
aft.aic(my_ml)
))
colnames(www)<- c("rs","kl","sd","dw","ml")
www  # table shows alternatives aft models' aic for each of the players

# Weibull AFT
est.wai<- function(df){
bbb<-sum(df$Runs==0)
df$Runs[which(df$Runs==0)]<- rep(0.001,bbb)
mw<- flexsurvreg(Surv(Runs, delta)~ Type+ Vanue + Chasing, data= df, dist = "weibull")
mw1<- survreg(Surv(Runs, delta)~ Type+ Vanue + Chasing, data= df, dist = "weibull")
est<-mw1$coefficients
df$res<- residuals(mw, type="coxsnell")
cmodel<- survfit(Surv(res, delta)~1, data=df)
plot(cmodel$time, cmodel$cumhaz, type="s", xlab="run",ylab="Cum.haz of residuals",
```

```
main="Cox Snell residuals")
abline(0,1,lty=2, col="red")
return( list(ew=est, summary(mw1)))
}

est.wai(my_player) # model estimates with cox snell residuals

#======== competing risk analysis
# preparing the data frame
summary(df_player)
df_player$Dismissal <- as.factor(df_player$Dismissal)
df_player$s.bowled<- ifelse(df_player$Dismissal=="bowled",1,0)
df_player$s.caught<- ifelse(df_player$Dismissal=="caught",2,0)
df_player$s.lbw<- ifelse(df_player$Dismissal=="lbw",3,0)
df_player$status<- df_player$s.bowled+df_player$s.caught+df_player$s.lbw
df_player.test<- subset(df_player, Type=="Test")
df_player.odi<- subset(df_player, Type=="ODI")
# for test
ci<- Cuminc(time = df_player.test$Runs, status = df_player.test$status)
#ci
names(ci)
bo<- ci$CI.1
ca<- ci$CI.2
lb<- ci$CI.3
times<- ci$time
plot(ca~times, type="s", col="purple", xlab="Runs",ylab="Cumulative Incidence
function", main="player Name in Test",
ylim=c(0,max(bo,ca,lb)), )
lines(bo~times, type="s", col="red")
lines(lb~times, type="s",col="orange4")
legend("topleft", legend = c("Bowled","Caught","lbw"),
       col=c("red","purple","orange4"),bty="n",lty=1, cex=0.6)
# similary for the ODI matches
```

The following code snippets are presented in a generalized manner and are applied separately to different datasets. The exploratory data analysis (EDA) codes are not included here. For the complete code, please refer to the provided link. https://github.com/SoumaryaBasak/Biostatiticians-insight-into-sports.git