Batting Performance Across Formats: Predicting ODI Strike Rates Using T20 Metrics*

Top 100 Players Excel in Both Formats, but Adaptability Remains Uniform Across Groups

Dhruv Gupta

December 3, 2024

This paper investigates the extent to which batting strike rates in T20 cricket predict performance in the ODI format, focusing on differences between top 100 players and their peers. Using ball-by-ball data from 17,944 cricket matches, we construct a linear model to estimate ODI strike rates based on T20 performance, player type, and their interaction. Results reveal a positive relationship between T20 and ODI strike rates for all players, with top 100 players exhibiting consistently higher ODI rates. Surprisingly, the interaction term is not significant, suggesting that adaptability between formats may not differ substantially between player groups. These findings provide a clearer understanding of performance transferability across cricket formats and highlight the nuanced factors influencing player success.

Table of contents

---

*Code and data are available at: https://github.com/dhruv5423/Top-100-Batters-T20-to-ODI-Strike-Rate-Mapping

1

1 Introduction

The sport of cricket has evolved greatly in recent times. Traditionally played over the course of 5 days, new, shorter formats are becoming increasingly popular due to their accessibility and faster pace of play. One Day International (ODI) and T20 formats are foremost among these formats - consisting of one innings of 50 overs and 20 overs respectively for either side. (International Cricket Council 2024) All three formats, Test, ODI, and T20 present different challenges and require different skill-sets. Lasting 5 days, test matches require endurance,

patience and strategy. In comparison, T20 cricket is much more exciting, with batsmen 'aiming for boundaries' as the teams try to score as many runs as possible in this short format. ODI cricket is a balance of the two - while relatively fast, strategy is key and slower play is more prevalent. (Hnrd 2024)

A batter's strike rate is a key performance indicator of their ability to score runs. It is defined as the 'runs per 100 balls' they face. (Sportmakers 2024) For example, if a batter scores 100 runs in 50 balls, their strike rate can be calculated to be 200. Given the differences in strategy between formats, it is common to expect differences in strike rates. Longer formats generally have lower strike rates, and shorter formats higher. But how well do strike rates translate across formats, particularly for top-performing players compared to the rest of field? This is of particular interest, not only in evaluating individual player adaptability, but in understanding how skills in one format can be extrapolated to another.

In this paper, we examine how T20 batting strike rates predict ODI performance, focusing on differences between top-performing players and their peers. We model ODI strike rates using T20 strike rates, player type (top 100 or non-top 100), and their interaction, analyzing ball-by-ball data from nearly 17,944 matches. Our findings reveal that while T20 strike rates positively correlate with ODI performance, the interaction term suggests that adaptability does not differ significantly between top 100 players and others. This challenges the assumption that top players are uniquely skilled at transferring their performance between formats.

These findings are important for talent evaluation and team selection strategies in cricket. By quantifying the relationship between T20 and ODI performance, this research provides clarity on the dynamics of skill transferability. The rest of paper is structured as follows: Section 2 contains information regarding data selection and sourcing, samples of the raw data, a description of the cleaning process, and an example of the cleaned analysis data. Section 3 contains the model used in this paper - the set up, the estimator, and the justification and assumptions used to design it. Section 4 provides a summary of the regression output, fitted values for the model, and graphical representation of the outcome variable. Section 5 is a discussion, analyzing the coefficients the model predicts, intepreting the graphs in Section 4, and using literature to draw broader conclusions regarding our findings. It also discusses key limitations in the model. Section 6 includes details on how to download the data and challenges with regards to sampling and observing cricket data.

2 Data

The statistical software R, (R Core Team 2023), is used throughout the paper to download, clean, and analyse the data. Additionally, many packages within the Tidyverse, (Wickham et al. 2019), Arrow, (Richardson, Hester, et al. 2023), and KableExtra, (Zhu 2023) are used.

2.1 Data Selection and Sourcing

Data used in this paper has been sourced from the reputable cricket database, CricSheet, (Rushe 2024). Created by Stephen Rushe, CricSheet is an open-source platform providing detail ball-by-ball data for 17,944 cricket matches. Contained in the database were 2,858 One Day International matches and 3,856 T20 International matches. This paper focuses on the male cricket matches, with 2,391 Male ODIs and 2,499 Male T20 match data downloaded as CSV files. This data set is particularly well-suited for this study because it ensures consistent measurement across matches, detailed coverage of player-level performance, and a large sample size, enabling robust statistical analysis of performance differences across formats.

Additional details on the data selection process and a step-by-step guide to access and download the data are provided in the appendix Section 6.

2.2 Measurement

The dataset used in this research is an organised and quantitative attempt to capture real-world cricketing phenomena, specifically player performance in the T20 and ODI formats. The data on CricSheet is derived from genuine ball-by-ball records of cricket matches, which are meticulously documented by official scorers throughout games. Each element in the dataset represents an observable event from a match, such as a ball delivered, runs scored, or wicket taken. This is a detailed, play-by-play breakdown of the game.

The process of converting these real-world cricketing occurrences into dataset entries begins with match officials and scorers recording the results of each ball. CricSheet standardised and digitised these logs into CSV files. Key variables including strike rate, player type, and total balls faced, were derived directly from these raw observations. For example, strike rate is calculated as the ratio of runs scored to balls faced, multiplied by 100—a straightforward transformation of raw data into a meaningful metric.

2.3 Raw Data

Below are samples from the raw data regarding ODI matches after all raw csv files of individual match data were combined. The data is split into logical groups for clarity, but are all part of the same data frame.

2.3.1 Match Information

Table 1 contains basic information about the match: match ID, season, start date, venue, innings, and ball. So, the first row would give match information of Match ID 1131279 at the ICC Academy in the 2017/2018 season, on the 1st of January 2023. The ball that the data captured is the 4th ball from the 44th over, in the first innings.

4

Table 1: Raw Data - Sample Match Information

Match Id	Season	Start Date	Venue	Innings	Ball
1131279	2017/18	2018-01-23	ICC Academy	1	44.4
1031673	2017	2017-09-27	Kennington Oval	2	26.6
656483	2014/15	2015-03-18	Sydney Cricket Ground	1	20.5

2.3.2 Batting and Bowling Teams

Table 2 is about the batting team, bowling team, and key players involved at the time of the delivery. The first row can be understood to show that the time of delivery 44.4 of Match 1131279 (correponding to row 1 in Table 1), the batting team is India, the bowling team is New Zealand, the Striker is V Kohli, the other batsman (or the non-striker) is SS Iyer, and the Bowler is TA Boult.

Table 2: Raw Data - Sample Batting and Bowling Teams Information

Batting Team	Bowling Team	Striker	Non Striker	Bowler
India	New Zealand	V Kohli	SS Iyer	TA Boult
Sri Lanka	South Africa	I Udana	CAK Rajitha	A Nortje
Sri Lanka	New Zealand	WU Tharanga	DPMD Jayawardene	SE Bond

2.3.3 Runs and Extras

Table 3 contains the runs scored, extras such as wides, no-balls, byes, leg-byes, and penalties. So for the same ball in Table 1 and Table 2, 1 run was scored by V Kohli. There were no extras, wides, no balls, byes, legbyes, or penalties.

Table 3: Raw Data - Sample Runs and Extras Information

Runs Off Bat	Extras	Wides	Noballs	Byes	Legbyes	Penalty
1	0	NA	NA	NA	NA	NA
0	0	NA	NA	NA	NA	NA
0	0	NA	NA	NA	NA	NA

2.3.4 Wicket Details

Table 4 contains information about wickets, including the type of dismissal and the players involved. For each of the 3 balls displayed in this sample there were no wickets or players

dismissed.

Table 4: Raw Data - Sample Wicket Details

Wicket Type	Player Dismissed	Other Wicket Type	Other Player Dismissed
		NA	NA
		NA	NA
		NA	NA

2.4 Cleaning

2.4.1 Desired Structure of Analysis Data

Refer to the "Sketches" section of the repository. There was an understanding of how the dataset needed to be transformed in order to get it ready for analysis through our model. he desired structure of the data aimed to include the following key variables: player name, player type, match type, strike rate, and total balls faced. Below is an example of a representation of the desired structure:

Table 5: Desired Structure of Analysis Data

Player	Player Type	Match Type	Strike Rate	Total Balls Faced
Player A	Top 100	T20	140	1200
Player A	Top 100	ODI	90	1200
Player B	Non-Top 100	T20	120	800
Player B	Non-Top 100	ODI	75	800

2.4.2 Filtering on Balls Faced

The first step after downloading raw data files and combining them into large T20 and ODI ball by ball data files was to filter players who had a large enough sample size across formats. That is to say, only those who meaningfully participated. Importantly we wanted to exclude players whose sample size of balls faced were too few, as these observations could skew the findings from the paper. Therefore, players were required to meet a minimum threshold of 200 balls faced in T20 matches and 500 balls faced in ODI matches. The combined balls faced in both formats needed to be above a threshold of 700. In doing so, we avoid noise from low-sample sizes. The library dplyr, (Wickham et al. 2023), was used to filter and aggregate the data.

2.4.3 Calculating Average Strike Rates, Identifying Top 100 Players

The 'mean' function was used to calculate the average T20 Strike Rate and average ODI strike rate. Missing or null values were exlcuded using 'na.rm = TRUE'.

Because part of are model is concerned with differences between top performing players and their peers, we obtain the Top 100 batsmen before y ranking them based on the average of their T20 and ODI strike rates. Here, the 'mutate' function was used to calculate ranks and assign each player as either 'Top 100', or 'Non-Top 100'.

2.4.4 Structuring the Final Dataset

The arrow package was used to save the cleaned data set as a Parquet file. (Richardson, Hester, et al. 2023) The final cleaned dataset now included the columns "Player", "Player Type", "Match Type", "Strike Rate", "Total Balls Faced". Table 6 contains a sample of the first 5 rows of the final analysis data.

Table 6: Sample of Analysis Data

Striker	T20 Strike Rate	ODI Strike Rate	Total Balls Faced	Player Type
CK Kapugedera	114.60	69.77	2746	Non-Top 100
MD Shanaka	116.76	87.55	2540	Non-Top 100
EJG Morgan	132.92	89.52	9605	Top 100
SA Edwards	122.68	90.66	2457	Non-Top 100
LRPL Taylor	118.49	80.94	11668	Non-Top 100

2.5 Summary of Analaysis Data

After getting the data ready for analysis, it may be useful to visualize the data before we run the model. Figure 1 displays the average strike rates for both match types, for both player groups. As expected, we find that Top 100 players have higher strike rates than Non-Top 100 players.
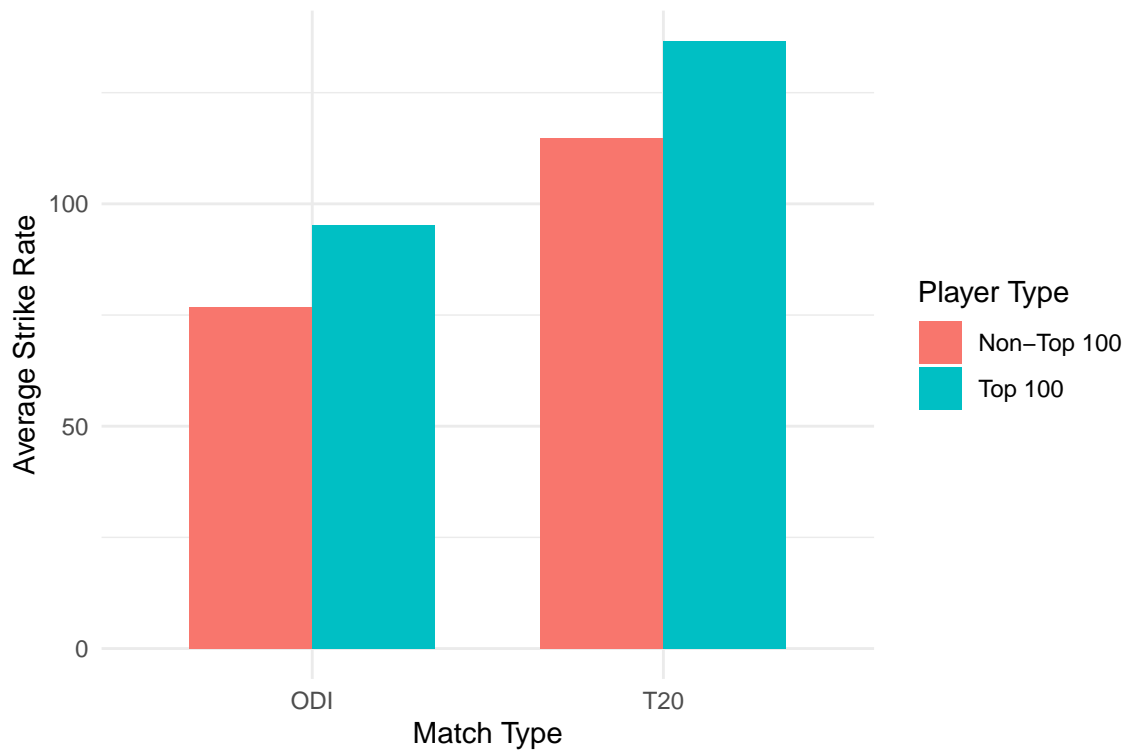
Figure 1: T20 vs. ODI Average Strike Rates by Player Group

Table 7 summarizes the average strike rates for each player type - numerically describing Figure 1.

Table 7: Average Strike Rates by Player Type

Player Type	T20 Strike Rate	ODI Strike Rate
Non-Top 100	114.7136	76.76375
Top 100	136.6294	95.16540

3 Model

3.1 Model Set-Up

In order to investigate how T20 strike rates translate into ODI Strike rates, we fit a linear regression model. Our outcome variable is Y_i, for which we attempt use 3 explanatory variables

to fit: T20 Strike Rate - X_{1i}, a dummy variable for player type (whether they are Top 100 or not) - X_{2i}, and an interaction term between the two, $X_{1i} \cdot X_{2i}$.

The mathematical representation of the model is as follows:

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 (X_{1i} \cdot X_{2i}) + \epsilon_i$$

Where:

- Y_i: The ODI strike rate of player i.

- X_{1i}: The T20 strike rate of player i.

- X_{2i}: A binary indicator of player type, where $X_{2i} = 1$ for "Top 100" players and $X_{2i} = 0$ for "Non-Top 100" players.

- $X_{1i} \cdot X_{2i}$: The interaction term between T20 strike rate and player type.

- β_0: The intercept, representing the predicted ODI strike rate for "Non-Top 100" players with a T20 strike rate of 0.

- β_1: The main effect of T20 strike rate, representing the change in ODI strike rate for every 1-unit increase in T20 strike rate among "Non-Top 100"

- β_2: The main effect of player type, representing the difference in ODI strike rate between "Top 100" and "Non-Top 100" players when T20 strike rate is 0.

- β_3: The interaction effect, representing how the relationship between T20 and ODI strike rates differs for "Top 100" players.

- ϵ_i: The residual error, assumed to be normally distributed with mean 0 and constant variance σ^2.

We fit the model using the `lm()` function in R, implemented in the **Base R** package.

Our response variable is Y, ODI Strike Rate. Our predictors are the continuous variable T20 Strike Rate, X_1, the binary variable Player Type, X_2, and the interaction term $X_1 \cdot X_2$.

3.2 Model Justification

We attempt to capture the relationship between T20 Strike Rates and ODI Strike Rates, while accounting for player type using a dummy variable. Further, the inclusion of the interaction term directly tackles the research question, allowing us to test whether the relationship between T20 and ODI Strike Rates differs between groups - demonstrating potential differences in their adaptability.

In designing this model, we make the following assumptions:

1. Linearity in Coefficients: We assume that there exists a linear relationship between T20 Strike Rates and ODI Strike Rates. The coefficient β_1 captures this. Additionally, the effect of player type, β_2, is also assumed to be linear. Furthermore, the coefficient of the interaction term, β_3, is also modeled linearly.

2. Residuals are Normally Distributed: This is to say that the variability of ODI Strike Rate around the predicted value we fit using the model should be normally distributed, around 0. Section 4 contains a histogram of the residuals - the model assumes normality in the form of a bell shaped curve.

3. Homoscedasticity: This assumption is built into the model to imply that the variance of the residuals, ϵ_i, is constant across independent variables. That is to say that the spread of the residuals should be homogenous regardless of whether we look at the predictor $\beta 1$ or $\beta 2$.

4. No Collinearlity: This assumption states that our predictor variables are not highly correlated with each other - this can make it difficult to interpret our coefficients are we are unsure regarding the interaction between the two, and if one is influencing the other.

4 Results

4.1 Regression Summary

Table 8 presents the coefficients from the linear regression model.

Table 8: Regression Coefficients for the Model

Term	Estimate	Standard Error	t-value	p-value
(Intercept)	54.8724138	7.2396259	7.5794543	0.0000000
t20_strike_rate	0.1908346	0.0628596	3.0358850	0.0026176
player_type_binary	19.3615909	14.5835919	1.3276284	0.1853522
t20_strike_rate:player_type_binary	-0.0376363	0.1117765	-0.3367103	0.7365805

4.2 Fitted Values

To understand the results of the model, we begin by examining the fitted values. These are the predicted ODI Strike Rates for each player based on the T20 Strike Rates, Player Type, and their interaction. The fitted values are computed using the following equation:

$$\hat{y} = \beta_0 + \beta_1 \cdot \text{T20 Strike Rate} + \beta_2 \cdot \text{Player Type Binary} + \beta_3 \cdot (\text{T20 Strike Rate} \times \text{Player Type Binary})$$

where:

- $\hat{y}$ is the predicted ODI Strike Rate,

- β_0 (Intercept) = 54.87,

- β_1 (T20 Strike Rate Coefficient) = 0.19,

- β_2 (Player Type Binary Coefficient) = 19.36,

- β_3 (Interaction Term Coefficient) = -0.038.

The fitted values provide insight into how well the model captures the observed data. Table 9 shows a sample of the observed ODI Strike Rates and their corresponding fitted values, grouped by Player Type.

Table 9: Sample of Fitted Values and Observed ODI Strike Rates

Striker	Player Type	T20 Strike Rate	ODI Strike Rate (Observed)	ODI Strike Rate (Fitted)	Residuals
Jatinder Singh	Non-Top 100	115.02	71.37	76.82	-5.45
CK Kapugedera	Non-Top 100	114.60	69.77	76.74	-6.97
KH Prajapati	Non-Top 100	103.95	73.72	74.71	-0.99
Aqib Ilyas	Non-Top 100	118.34	72.27	77.46	-5.19
H Klaasen	Top 100	137.64	113.90	95.32	18.58

4.3 Graphical Representation

4.3.1 Predicted vs. Observed ODI Strike Rates

Figure 2 evaluates the accuracy of the model by comparing the predicted ODI Strike Rates to the actual observed values.
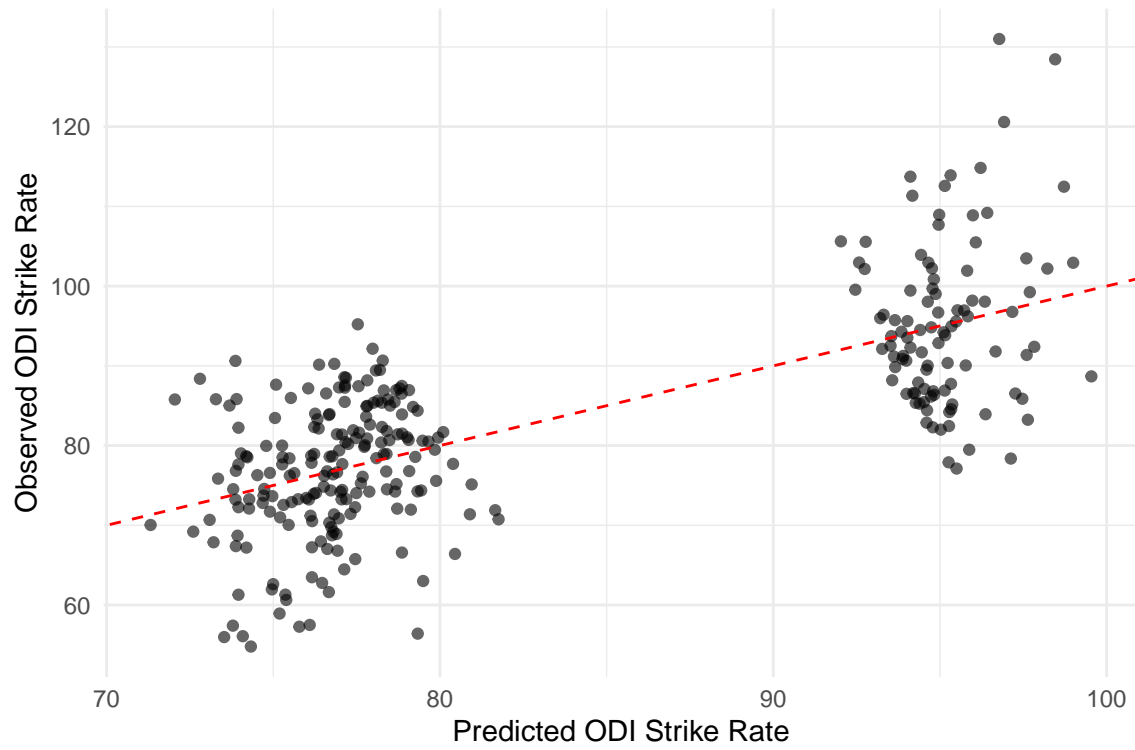
Figure 2: Predicted vs. Observed ODI Strike Rates

4.3.2 Residuals Distribution

Figure 3 assesses the assumption of normally distributed residuals by visualizing the spread of residuals around zero, in the form of a histogram.
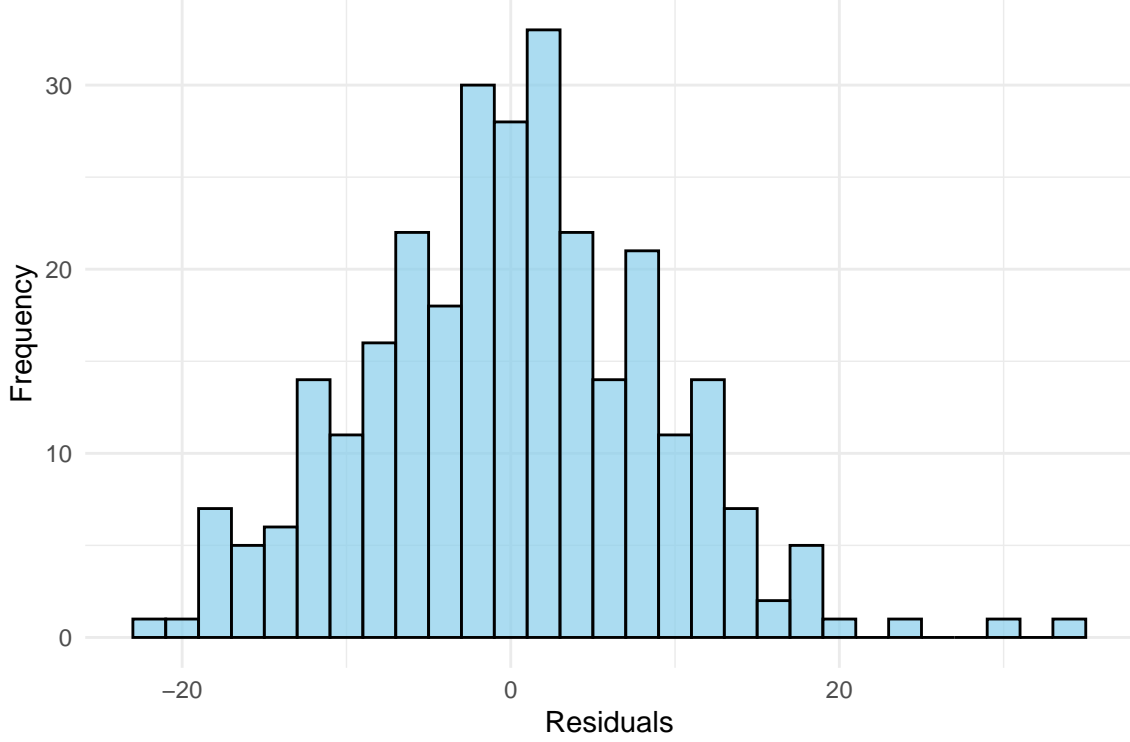

Figure 3: Distribution of Residuals

5 Discussion

5.1 Intepretation of Model Coefficients

The linear regression model estimates the relationship between T20 and ODI strike rates while accounting for player type. From Table 8, we can use the coefficients from our model to generate the following:

$$\hat{y} = 54.87 + 0.19 \cdot \text{T20 Strike Rate} + 19.36 \cdot \text{Player Type Binary} - 0.037 \cdot (\text{T20 Strike Rate} \times \text{Player Type Binary})$$

The intercept, β_0, indicates that the expected strike rate for a non-top 100 player in an ODI match is 54.87. That is to say, if faced with 100 balls, the average non-top 100 player scores almost 55 runs. The coefficient $\beta_1 = 0.19$ suggests that, on average, a one-unit increase in T20 strike rate is associated with a 0.19-unit increase in ODI strike rate for non-top players, statistically significant at p = 0.002. This positive relationship highlights that higher T20 strike rates generally correspond to better performance in ODIs for these players.

Interestingly, we observe that our dummy variable, 'Player Type Binary', has a coefficient $\beta_2 = 19.36$. This reflects the average difference in ODI strike rates between top and non-top players, controlling for T20 strike rate. Although this difference is not statistically significant (p = 0.185), it provides an approximate measure of the expected performance gap. As expected, higher skilled players tend to score runs at a higher rate.

The interaction term has a coefficient β_3 = -0.037, which is not statistically significant (p = 0.737). From the predictions of the models we can reasonably say that there is no evidence that the relationship between T20 strike rates and ODI strike rates differs between top performing players and others. According to our model, what is a better predictor of success is the skill level that top performing batsmen exhibit.

5.2 Intepretation of Graphs

Figure 2 shows that most data points are clustered around the 45 degree line. This demonstrates that the model captures much of the variability in ODI strike rates. There are however, some outliers that might indicate endogeneity or other confounding variables. Figure 3 show that the value of our residuals are roughly normally distributed, and centered around zero. This is in line with Assumption 2 in Section 3.1, and we can be reasonably confident that our model predictions are unbiased on average. However, there are some outliers towards the tails of the curve, potentially indicating the model's limitation in capturing these extreme values.

5.3 Broader Context

The analysis done in this paper highlights the transferability between T20 performance and ODI performance. With a high degree of confidence we can say that higher T20 strike rates will generally result in high ODI strike rates, for all players - regardless of whether they are a top performing player or not. As we expect, top players exhibit a much higher strike rate than non top 100 players in ODI matches - but the lack of a significant interaction term tells us that our hypothesis was misguided. The model tells us that unintuitively, top players may not show any additional 'adaptability' separating them from their peers. Additionally, the low value of the coefficient β_1 indicates that it is quite difficult to increase ODI strike rate. For a given batsman's strike rate to increase by 1 run per 100 balls, the model predicts that they would need more than 5 runs per 100 balls in the T20 format.

The findings of this study align with existing literature that emphasizes the interconnectedness of cricket formats and the transferability of batting skills. By highlighting the positive correlation between T20 and ODI strike rates, this research contributes to a deeper understanding of how modern batting techniques and strategies are evolving in response to the demands of different formats.

5.4 Limitations

There are a fair amount limitations that are present in this model. First and foremost, the model uses strike rate as the sole predictor of batting performance - a top 100 player was defined purely off of strike rate. This approach ignores other critical aspects of batting performance, such as average runs scored, consistency, and the ability to perform under pressure. Players who may not have exceptionally high strike rates but consistently score runs over long periods could also be considered top performers in broader contexts. Thus, relying solely on strike rates could lead to an incomplete characterization of player quality.

Furthermore, the model assumes linearity in the relationship between T20 and ODI strike rates, which might oversimplify the actual dynamics. Batting performance could exhibit non-linear trends, especially as players adapt their approach to match contexts or as they age and gain experience. These dynamics are not captured in the current framework

There are more factors to consider. Environmntal and equipment constraints, such as pitch conditions, ball type, and equipment modifications, also play crucial roles in batting performance. These factors can affect a player's ability to adapt their skills across different formats and conditions. (**sage?**) A better model would likely incorporate these influences.

6 Appendix

This section contains additional data details.

6.1 How to Download the Data

To replicate the analysis, follow these steps to download the datasets:

1. Navigate to CricSheet: Visit the CricSheet website.
2. Locate Match Type Data:
3. Click on the relevant format (e.g., ODI, T20).
4. Download the ball-by-ball CSV files for each format under "Men's Cricket" section.
5. Extract the Files:
6. The downloaded files will be in ZIP format.
7. Extract all files into a folder labeled according to the match format (e.g., ODI_matches/, T20_matches/).

6.2 Sampling and Observational Data:

The analysis is based on data that is observational in nature - we did not conduct experiments or control any variables. Observational data in sports analytics often arises from recorded performance metrics rather than randomized experiments. While this approach provides real-world information, it comes with limitations such as unmeasured confounding variables and selection bias.

Specific to cricket there may be challenges in samping:

- Inconsistent Match Representation: Players may participate in varying numbers of matches across formats, leading to unbalanced observations.
- Player Inclusion Criteria: A threshold for balls faced was used to filter players with sufficient data (e.g., 200 balls for T20, 500 for ODI). This threshold minimizes noise but introduces potential bias by excluding players with fewer matches.
- Team and Venue Effects: Factors like team strength or venue conditions could influence performance, but these are not explicitly modeled.

6.3 Idealized Methodology

To address these challenges in future work, an idealized methodology could involve:

- Stratified Sampling: Stratify players by their participation levels (e.g., high, medium, low match participation). This also ensures balanced representation across player groups for analysis.

- Better Simulated Data: Simulate player performances under controlled conditions to test hypotheses. For example, simulate strike rates by varying contextual factors like opposition quality or pitch conditions.

- Incorporating Hierarchical Models:Use hierarchical (mixed-effects) models to account for team and venue effects. This approach would allow for varying intercepts/slopes across teams or venues, reducing bias from unmeasured variables.

Citations

Hnrd. 2024. "Cricket Analysis and Insights." https://www.behind.hnrd.co/blog/6jmr1026fpsj9mdn2orestodtiegcl.

International Cricket Council. 2024. "The Three Formats of Cricket." https://www.icc-cricket.com/about/cricket/game-formats/the-three-formats.

R Core Team. 2023. *R: A Language and Environment for Statistical Computing.* Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/.

Richardson, Neal, Jim Hester, et al. 2023. *Arrow: Integration to 'Apache Arrow' Library.* https://CRAN.R-project.org/package=arrow.

Rushe, Stephen. 2024. "CricSheet: Ball-by-Ball Cricket Data for Male ODI and T20 Matches." https://cricsheet.org/downloads/. https://cricsheet.org/downloads/.

Sportmakers. 2024. "What Is the Strike Rate in Cricket?" https://www.sportmakers.co.uk/cricket/what-is-the-strike-rate-in-cricket/.

Wickham, Hadley, Mara Averick, Jennifer Bryan, Winston Chang, Lucy D'Agostino McGowan, Romain François, Garrett Grolemund, et al. 2019. "Welcome to the tidyverse." *Journal of Open Source Software* 4 (43): 1686. https://doi.org/10.21105/joss.01686.

Wickham, Hadley, Romain François, Lionel Henry, and Kirill Müller. 2023. *Dplyr: A Grammar of Data Manipulation.* https://CRAN.R-project.org/package=dplyr.

Zhu, Hao. 2023. *kableExtra: Construct Complex Table with 'Kable' and Pipe Syntax.* https://CRAN.R-project.org/package=kableExtra.