Analyzing Factors Affecting Pakistan Cricket Team's Performance (1952-2024)*

Opponent and Coin Toss Play Key Role in Deciding Match Outcome

Muhammad Abdullah Motasim

December 11, 2024

This paper analyzes Pakistan's cricket performance from 1952 to 2024, using a Bayesian Logistic Regression model to examine factors influencing match outcomes. We find that opponent strength and toss outcomes are significant predictors of Pakistan's success. These findings provide information for coaches, analysts, and administrators, offering strategies to improve performance under varying conditions. Ultimately, our analysis contributes to a deeper understanding of the factors that drive Pakistan's cricket success, with implications for teams worldwide.

Table of contents

1 Introduction **2**
- 1.1 Estimand . . . 3

2 Data **3**
- 2.1 Raw Data . . . 3
- 2.2 Data Cleaning . . . 4
- 2.3 Measurement . . . 5
- 2.4 Variable Analysis . . . 5

3 Model **6**
- 3.1 Model Overview . . . 6
- 3.2 Model Equation . . . 6
- 3.3 Model Priors . . . 8
- 3.4 Model Justification . . . 9

*Code and data are available at: https://github.com/abdullah-motasim/Analyzing-Pakistans-Cricket-Data.

4 Results **9**

5 Discussion **16**
5.1 Analysis Overview 16
5.2 Effect of Opponent 17
5.3 Matches Won Over the years 17
5.4 Weaknesses and Limitations: 18
5.5 Next Steps 19

Appendix **20**

A Surveys, Sampling, and Observational Data **20**
A.1 Data Collection 20
A.2 Annual Player Performance 21
A.3 Simulation Study 21
A.4 Bias-Adjustment Techniques 22

B Model details **22**
B.1 Model assumptions 22
B.2 Model limitations 22
B.3 Model validation 23
B.4 Model diagnostics 24

References **25**

1 Introduction

Pakistan's cricket team holds a storied legacy, marked by remarkable highs and challenging lows. From their 1992 World Cup triumph to fluctuations in form over the decades, analyzing trends in performance offers information into their journey and areas for strategic improvement. Despite the wealth of player statistics and match highlights available, there has been a lack of systematic statistical analysis that evaluates long-term performance trends. This paper fills that gap by analyzing data from 1952 to 2024 to uncover the key factors that contribute to Pakistan's success in international cricket.

We employ a Bayesian Logistic Regression model to analyze match outcomes, considering variables like year, opponent, and toss outcomes. Our findings reveal several important observations. First, the strength of the opposing team emerged as a significant factor influencing Pakistan's chances of winning. We also found that toss outcomes had a noteworthy impact on match results. This paper highlights the importance of both internal factors (such as team performance) and external factors (such as opponent strength) in shaping match outcomes.

These observations provide strategic guidance for coaches, analysts, and cricket administrators, not only in Pakistan but also around the world, on how to prepare and perform under varying conditions.

The structure of this paper is as follows: Section 2 discusses the data types included in the raw data, the cleaning process for the data, and the reason for selecting the data set we did. Section 3 discusses model specification and justification for a Bayesian Logistic Regression model. Section 4 analyzes the trends and correlations between different variables utilizing tabular and graphical means. Section 5 discusses the results of Section 4 going into detail on what the simulation results can tell us about Pakistan cricket performance, as well as assumptions and limitations in data.

1.1 Estimand

The estimand is the probability that Pakistan wins a cricket match, given the opponent, toss outcome, and year.

2 Data

The data for this study were obtained from the cricketdata R package (Hyndman et al. 2024) and all analysis was performed using `R` (R Core Team 2023) alongside the following packages: `tibble` (Müller and Wickham 2023), `readr` (Wickham, Hester, and Bryan 2024), `arrow` (Richardson et al. 2024), `tidyverse` (Wickham et al. 2019), `dplyr` (Wickham et al. 2023), `testthat` (Wickham 2011), `caret` (Kuhn and Max 2008), `rstanarm` (Vehtari et al. 2022), `tidyr` (Wickham, Vaughan, and Girlich 2024), `modelsummary` (Arel-Bundock 2022), `knitr` (Xie 2023), `kableExtra` (Zhu 2024) and `ggplot2` (Wickham 2016).

As mentioned before the raw data was collected from the cricketdata library which contains data on international and other major cricket matches sourced from ESPN Cricinfo (ESPN Cricinfo 2024) and Cricsheet (Cricsheet 2024). It contains details on the individual match year, teams, result, and toss winner. Additionally, the data set features career statistics for individual players, including total runs, total balls faced, total wickets, and more. This data set was chosen for its ease of implementation within R, as it is available as a package, and for its ability to scrape data from multiple reputable cricket sources, such as ESPN, thereby enhancing the overall reliability of the data set.

2.1 Raw Data

The cricketdata library offers a variety of functions that provide data sets on different cricket statistics. However, for our analysis, we focus on two key areas: individual match statistics, to

determine which matches Pakistan played in and how they performed, and player career statistics, to assess how the quality of players has evolved over time and its impact on Pakistan's overall performance. To extract the relevant data, we primarily used two functions from the library: fetch_cricinfo, which retrieves individual career performance data, and fetch_cricsheet, which provides detailed match information.

2.2 Data Cleaning

The raw match data initially contained 820 observations across 25 variables, each detailing a match between two countries. The data cleaning process involved selecting only matches in which Pakistan played, and retaining only the variables relevant to our analysis. These variables include:

- `Team1` - The first team that participated in the match.
- `Team2` - The second team that participated in the match.
- `Date` - The date of the match (eg. 2008-01-02)
- `Winner` - The team that won the match between Team1 and Team2 (NA if the match was a draw).
- `Winner_wickets` - The number of wickets the winning team took (NA if the match was a draw).
- `Winner_runs` - The number of runs the winning team scored (NA if the match was a draw).
- `Toss_winner` - The team that won the coin toss.
- `Toss_decision` - The decision made by the toss winner, indicating whether they chose to field or bowl.

The raw player career data was broken into batting, bowling, and fielding categories with some players appearing in multiple data sets, thus the cleaning process involved combining the 3 data sets and ensuring proper handling of one player in each set. After this, only Pakistani cricket players were selected and variables which were important to use were retained. These variables are:

- `Player` - Name of the player.
- `Start` - The year in which the player began their career
- `End` - The year in which the player's career concluded.
- `Matches` - The total number of matches the player participated in throughout their career.
- `Bat_innings` - The total number of batting innings the player participated in.
- `Bowl_innings` - The total number of bowling innings the player participated in.
- `Field_innings` - The total number of fielding innings the player participated in.
- `Bat_runs` - The total number of runs the player scored while batting.
- `Not_puts` - The number of times the player remained not out while batting.

- `Bat_average` - The player's batting average, calculated as the total number of runs scored divided by the number of times they were dismissed.
- `Bowl-runs` - The total number of runs conceded by the player while bowling. This statistic reflects how many runs the player allowed the opposition to score during their bowling spells.
- `Wickets` - The total number of wickets taken by the player in their career.
- `Economy` - The average number of runs the player concedes per over bowled. It is calculated by dividing the total number of runs given by the total number of overs bowled. A lower economy rate typically indicates a bowler who is more effective at restricting the opposition's scoring.
- `Dismissals` - The total number of dismissals the player has contributed to as a fielder or wicketkeeper.

After applying these cleaning procedures, we were left with 151 observations across 7 variables for the matches Pakistan has played in, and 344 observations across 14 variables for the career information of individual Pakistani cricketers.

2.3 Measurement

The cricketdata package extracts match information from official sources, including the ESPN and Cricsheet databases, scorecards, and records maintained by cricket boards. These sources provide detailed records of each international match played by Pakistan and other cricketing teams worldwide. These matches represent the real-world phenomena we aim to analyze.

This real-world data is transformed into structured entries in our dataset through the manual recording of match information by officials during the game. Key details, such as match ID, participating teams, venue, date, event, toss winner/decision, player of the match, umpires, match winner, and more, are carefully documented. Once recorded, this information is made available for download on official websites. The datasets are then retrieved and imported into R using the cricketdata package.

This measurement approach is highly reliable, as it relies on official cricketing records. However, it also highlights the limitations inherent in working with aggregated datasets. For example, factors that can influence match outcomes—such as team morale or crowd effects—remain unquantified.

2.4 Variable Analysis

Figure 1 show key match statistics for Pakistan. The data shows that Pakistan frequently faces England, Sri Lanka, and Australia, with a particularly poor record against Australia, including several draws.

Regarding coin toss outcomes, Sri Lanka and South Africa appear to have won a substantial number of tosses. However, it's important to note that while the graphs for match outcomes and tosses won may give the impression that Pakistan was victorious in most of these events, we must keep in mind that the data has been filtered to include only Pakistan's matches. This means Pakistan appears more frequently in the data set compared to other teams, which accounts for the higher statistics shown.

The two graphs showing the number of matches and tosses that Pakistan won or lost inform us that Pakistan's win-loss ratio is approximately 0.5. This indicates that Pakistan has lost about twice as many matches as it has won. Additionally, the coin toss win ratio is close to even, reflecting a near 50/50 chance of Pakistan winning the toss.

Finally, the graph illustrates the number of matches Pakistan played each year has a peak in 2016. It then gradually declines over the subsequent years, culminating in a sharp drop in the most recent year—comparable to the period during the COVID-19 pandemic.

3 Model

I used a Bayesian Logistic Regression model to estimate the probability of Pakistan winning a cricket match based on several key factors. Logistic regression is a statistical technique used for binary classification, where the goal is to predict the probability of a categorical outcome. In this case, the outcome is whether Pakistan wins or loses a match.

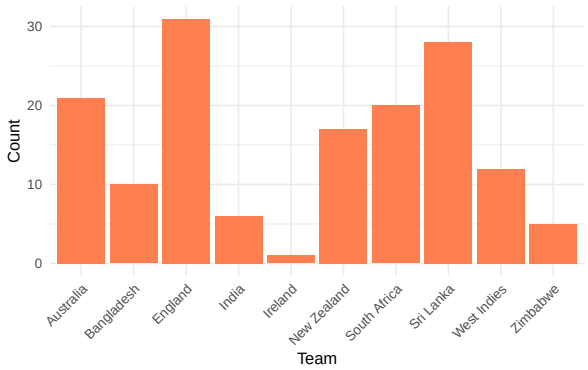
3.1 Model Overview

The model was trained in R utilizing the rstanarm package and it incorporates three independent variables that may influence match outcomes: year, opponent, and toss winner. The dependent variable is match_outcome, where 1 indicates a win for Pakistan and 0 indicates a loss. This setup allows us to examine the influence of key variables on Pakistan's performance.

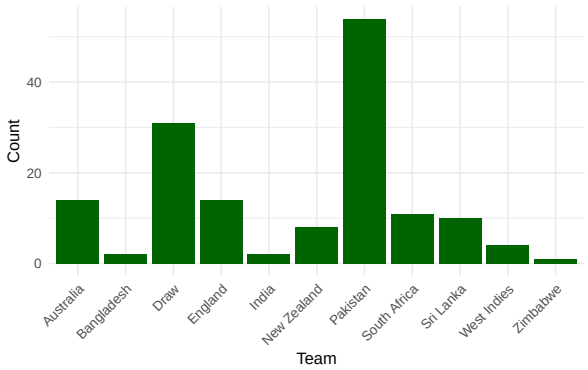
By incorporating these variables, the model explores how factors such as the year of the match, the opposing team, and whether Pakistan won the toss (which influences the decision to bat or bowl) affect Pakistan's likelihood of winning. This approach helps uncover trends in Pakistan's performance over time and under varying conditions, providing intuition into factors that may inform future match strategies.

3.2 Model Equation

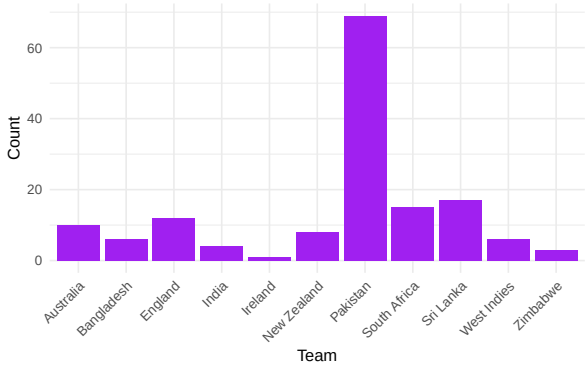
The model I will be using can be described using the following equation:



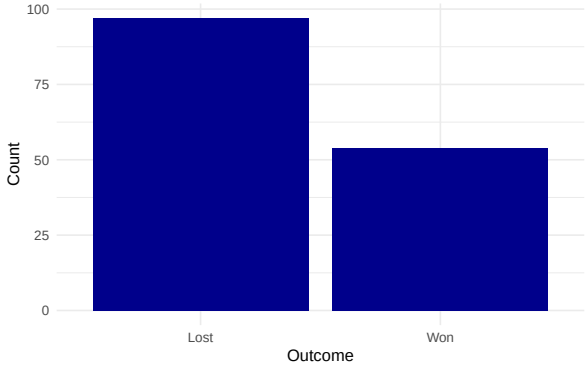
(a) Pakistan Opponent Team



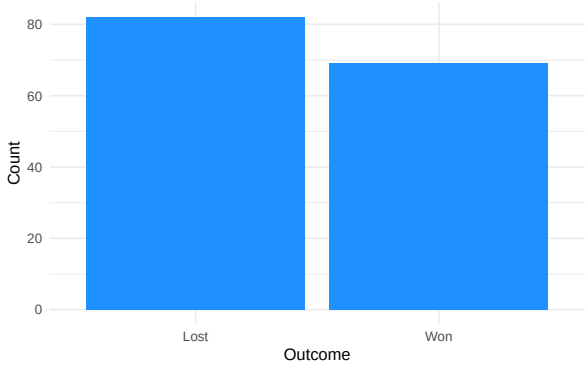
(b) Matches Won by Each Team



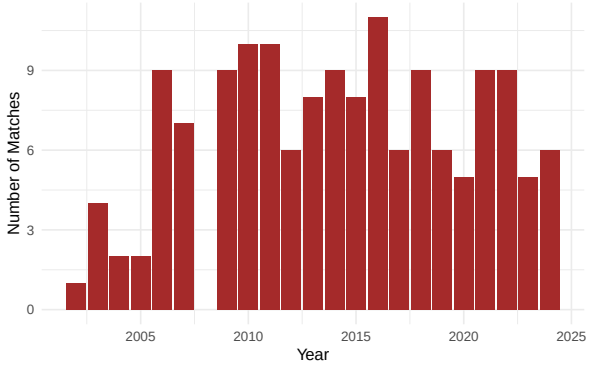
(c) Tosses Won by each Team



(d) Matches Won by Pakistan



(e) Tosses Won by Pakistan



(f) Pakistan's Matches Played per Year

Figure 1: Counts for match variables of interst

$$\log\left(\frac{\hat{p}}{1-\hat{p}}\right) = \beta_0 + \beta_1 \times \text{year} + \beta_2 \times \text{opponent} + \beta_3 \times \text{toss winner} \tag{1}$$

$$\begin{aligned}
\beta_0 &\sim \text{Normal}(0, 2.5) \\
\beta_1 &\sim \text{Normal}(0, 2.5) \\
\beta_2 &\sim \text{Normal}(0, 2.5) \\
\beta_3 &\sim \text{Normal}(0, 2.5)
\end{aligned}$$

where,

- $\hat{p}$ is the predicted probability of Pakistan winning a match.
- β_0 is the intercept (the log-odds of winning when all predictors are at their reference level).
- β_1 is the coefficient for the year variable, representing how match outcomes are affected by the year of the match.
- β_2 are the coefficients for each level of the categorical variable team2 (the opponent), with the reference category excluded. This allows the model to compare the odds of winning against different opponents.
- β_3 is the coefficient for the toss_win variable (whether Pakistan won the toss).

3.3 Model Priors

The priors used in the model represent our assumptions about the parameters before observing the data. These priors are chosen based on reasonable beliefs about the likely values of the coefficients and are specified as follows:

- `Prior for the Coefficients (year, team2, toss_win)`: Each coefficient is assigned a Normal prior with a mean of 0 and a standard deviation of 2.5, reflecting the assumption that the effects of the predictors are centered around zero with a reasonable amount of uncertainty.
- `Prior for the Intercept`: A Normal prior with a mean of 0 and standard deviation of 2.5 is used, representing the baseline log-odds of Pakistan winning a match when all predictors are zero.
- `Prior for the Auxiliary Parameter (prior_aux)`: This represents the dispersion of the logistic regression model and is assigned an Exponential prior with a rate of 1, typical for Bayesian Logistic Regression. It reflects uncertainty in outcome variance, with higher values suggesting more variability.

The Bayesian approach incorporates prior knowledge (through these priors) and updates it with the data from the training set, enabling the model to estimate relationships while accounting for uncertainty in the coefficients.

3.4 Model Justification

A decision tree model was also considered, which could capture non-linear relationships and interactions between predictors. However, decision trees are more prone to overfitting and can be less interpretable than logistic regression, especially in cases with categorical variables like team2.

An alternative approach considered was a frequentist logistic regression model. This model would not incorporate prior information and would instead focus purely on the likelihood of the data. However, the Bayesian approach was chosen due to its ability to incorporate prior knowledge and handle uncertainty more effectively, especially in the context of small sample sizes and the complexity of cricket data.

This model provides a useful framework for estimating Pakistan's likelihood of winning a match based on key factors. By using a Bayesian Logistic Regression approach, the model incorporates uncertainty in the estimates, giving us a clear interpretation of the effects of each predictor, and is flexible enough to adapt to new data as it becomes available. However, it should be acknowledged that there are limitations regarding omitted variables and potential model assumptions that could be addressed in future iterations.

This is an overview of the model. More details such as assumptions, limitations, convergence, and validation can be found in Section B

4 Results

Our results are summarized in Table 1 and Figure 2 presenting the predicted values of the coefficients. The figure also includes the 95% confidence intervals for the predictors, as estimated by our Bayesian model. A 95% confidence interval indicates that there is a 95% probability that the true value of the parameter lies within the specified range, given the observed data. In the graph, each point represents the predicted value, and the horizontal line extending from each point shows the 95% confidence interval.

Variables to the right of 0 have a positive correlation with the outcome, meaning that increasing these values increases the probability of Pakistan winning. Conversely, variables to the left of 0 exhibit a negative correlation, meaning that increasing these variables reduces the probability of Pakistan winning a match.

Finally, it is important to note that Figure 2 is plotted without the intercept term and the team2Ireland term. As shown in Table 1, these two values have much larger magnitudes compared to the other predictors. If included in the plot, their scale would be so large that the smaller variations in the other coefficients would be obscured.

The regression model results are as follows:

Table 1: Coefficient estimates and standard errors

	Model
(Intercept)	−55.46
	(79.33)
year	0.03
	(0.04)
team2Bangladesh	3.58
	(1.35)
team2England	1.11
	(0.74)
team2India	0.13
	(1.48)
team2Ireland	19.92
	(15.92)
team2New Zealand	0.68
	(0.83)
team2South Africa	0.23
	(0.93)
team2Sri Lanka	1.05
	(0.72)
team2West Indies	2.20
	(0.86)
team2Zimbabwe	2.47
	(1.51)
toss_win	0.59
	(0.43)

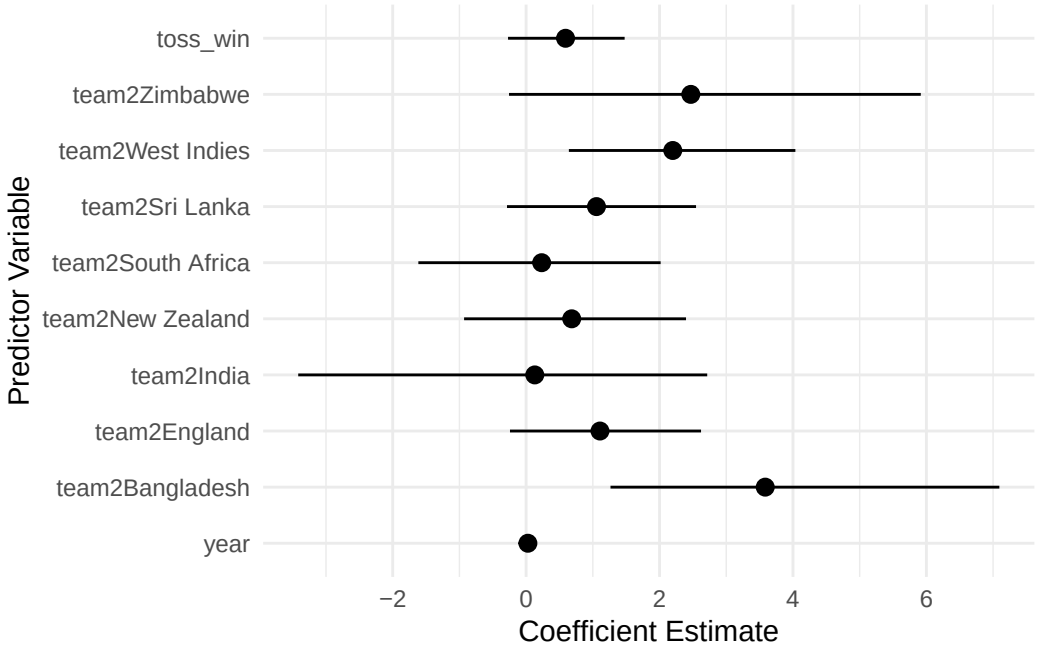


Figure 2: Coefficient estimates and 95% confidence intervals

- **Intercept**: The intercept of -55.94 (standard error = 80.63) represents the log-odds of Pakistan winning when all other variables are at their reference categories. While large in magnitude, it is not directly interpretable. The negative value suggests a very low baseline probability of winning without other influencing factors.
- **Year**: The coefficient for year is 0.03 (standard error = 0.04), indicating a very slight increase in the probability of Pakistan winning over time. The small magnitude and narrow standard error suggest the effect is minimal and the estimate is fairly precise.
- **Opponent**: The coefficients for opponents show how Pakistan's likelihood of winning changes against different teams, relative to the reference team:
  - `team2Bangladesh` has a coefficient of 3.62 (standard error = 1.31), which indicates a significant positive association with Pakistan's success. Matches against Bangladesh appear to increase Pakistan's chances of winning.
  - `team2Sri Lanka` shows a positive coefficient of 1.09 (standard error = 0.72), suggesting that matches against Sri Lanka are associated with a higher probability of Pakistan winning compared to the reference team, though the effect is less pronounced than for Bangladesh.
  - The coefficients for other teams, such as `team2India` (0.24, standard error = 1.44) and `team2South Africa` (0.29, standard error = 0.90), are much smaller, indicating

weaker associations with match success.

  - `team2Ireland` has the largest coefficient (19.77, standard error = 15.67), indicating a very high probability of success when playing against Ireland. However, this coefficient also has a large standard error, meaning there is significant uncertainty around this estimate.

- `Toss Win`: The coefficient for toss_win is 0.60 (standard error = 0.45), suggesting that winning the toss slightly increases Pakistan's likelihood of winning the match. This result supports the common understanding that the team winning the toss often gains an advantage in deciding whether to bat or bowl first.

Figure 3 illustrates the distribution of match outcomes (wins, losses, and draws) for Pakistan against eight different cricket teams: Sri Lanka, New Zealand, England, Australia, South Africa, West Indies, Zimbabwe, and Bangladesh. These results provide an overview of Pakistan's historical performance against each team and highlight patterns in their match outcomes.

- `Sri Lanka`: Pakistan's matches against Sri Lanka are fairly balanced, with a slightly higher frequency of losses compared to wins. Draws are the least frequent outcome, suggesting a competitive dynamic between the two teams.

- `New Zealand`: Matches against New Zealand reveal a similar trend, with losses being more frequent than wins. The proportion of draws remains relatively low, indicating fewer instances of matches being evenly contested.

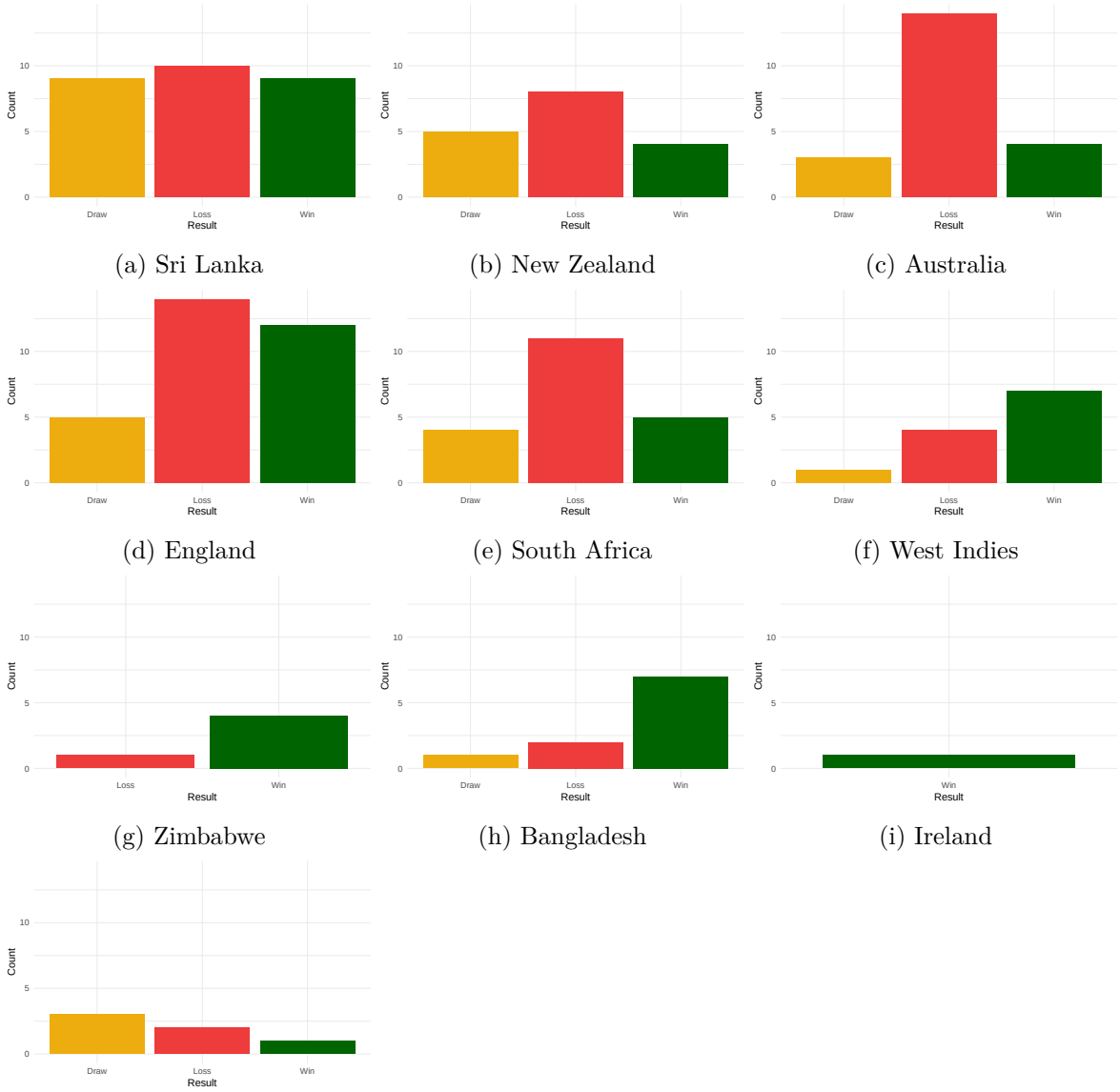
- `England`: Pakistan has a notably higher number of losses against England, with draws constituting the smallest category. This indicates that England has historically been a challenging opponent for Pakistan.

- `Australia`: A significant disparity is observed in matches against Australia, where losses outnumber wins by a considerable margin. This trend underscores the dominance of Australia in these encounters.

- `South Africa`: Matches against South Africa also show a higher proportion of losses compared to wins. However, the gap between these two outcomes is less pronounced compared to matches against Australia.

- `West Indies`: Pakistan exhibits a stronger performance against West Indies, with wins slightly outnumbering losses. Draws remain the least frequent outcome, emphasizing decisive results in these encounters.

- `Zimbabwe`: Against Zimbabwe, Pakistan demonstrates a clear dominance, with wins significantly outnumbering losses. Draws are nearly negligible, highlighting Pakistan's consistent success in these matches.



(a) Sri Lanka (b) New Zealand (c) Australia

(d) England (e) South Africa (f) West Indies

(g) Zimbabwe (h) Bangladesh (i) Ireland

(j) India

Figure 3: Pakistan's wins and losses against all teams

- `Bangladesh`: Similar to Zimbabwe, Pakistan has a markedly higher number of wins against Bangladesh, with losses and draws being infrequent. This dominance underscores Pakistan's strong track record against this opponent.

Overall, these graphs provide useful information into Pakistan's historical performance trends across different opponents. While some teams, such as Australia and England, have posed significant challenges, others, like Bangladesh and Zimbabwe, have been more favorable matchups for Pakistan. These findings align with the regression model coefficients presented earlier, where the opponent variable was a significant predictor of Pakistan's success in matches. For instance, the high win rates against Bangladesh and Zimbabwe correspond to their positive regression coefficients, while the challenging outcomes against teams like Australia reflect negative or smaller coefficients.

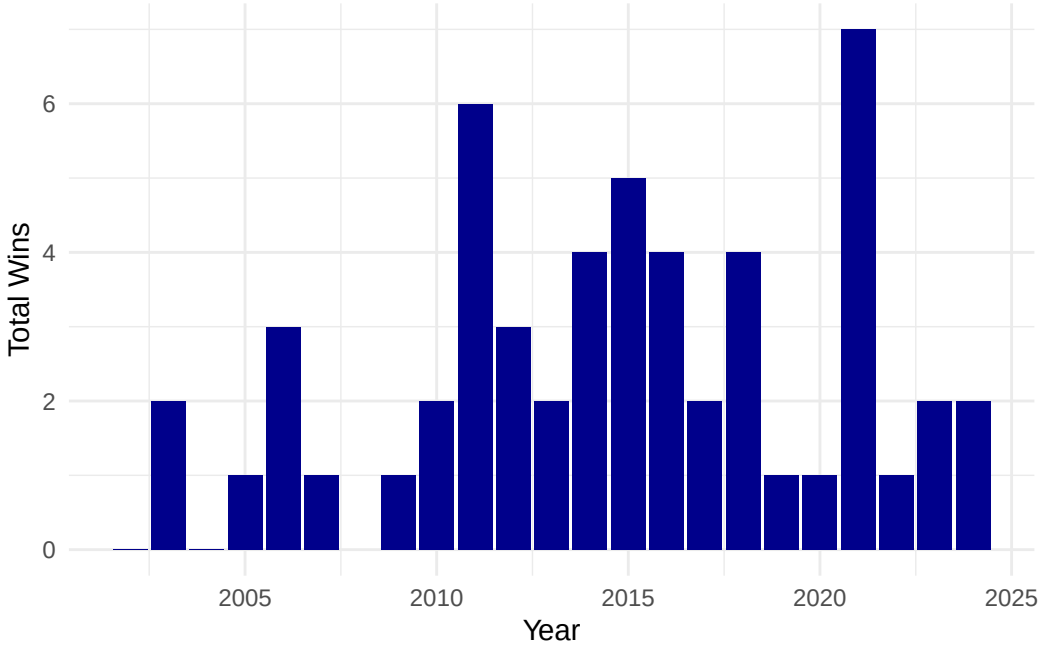


Figure 4: Number of Pakistan's wins over the years

Figure 4 displays the number of wins Pakistan has achieved each year. From the graph, we observe that Pakistan had a modest performance from 2002 to 2010, with only 2-4 wins per year, followed by a sharp increase in 2011. After this surge, the team maintained an average of 4 wins per year, with some fluctuations. However, in 2019, there was a notable decline, with just one win recorded. Since this drop, Pakistan has struggled to reach their previous high of 4 wins per year, and the current average stands at approximately 2 wins per year.

Figure 5 presents a scatter plot showing the years in which players retired and the number of runs they scored throughout their careers, with each dot representing an individual Pakistani

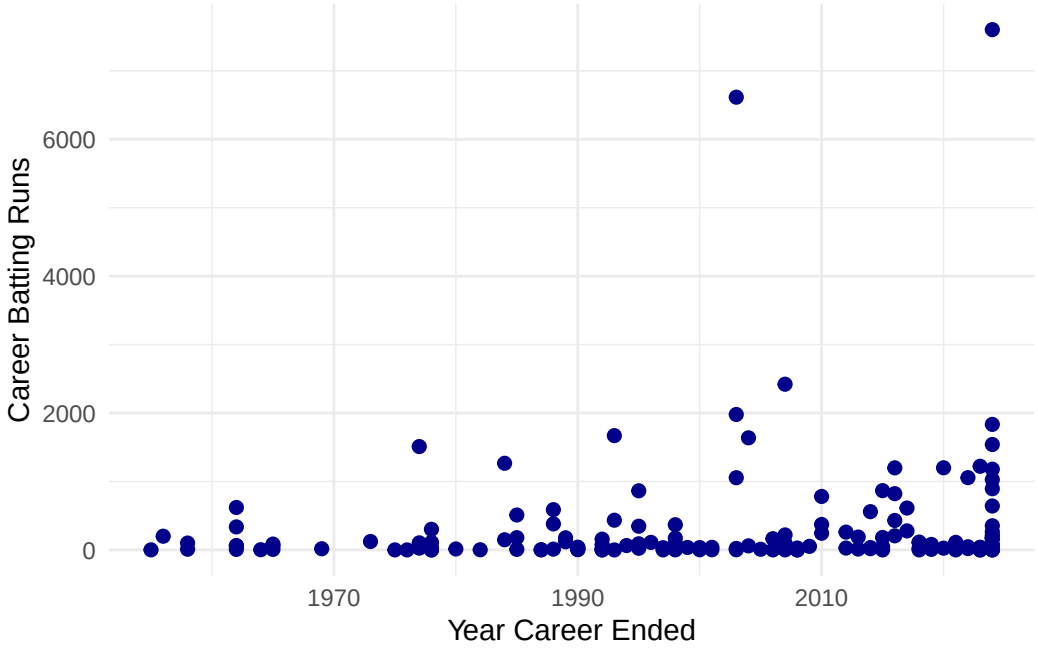


Figure 5: Career batting runs for each player by year of retirement

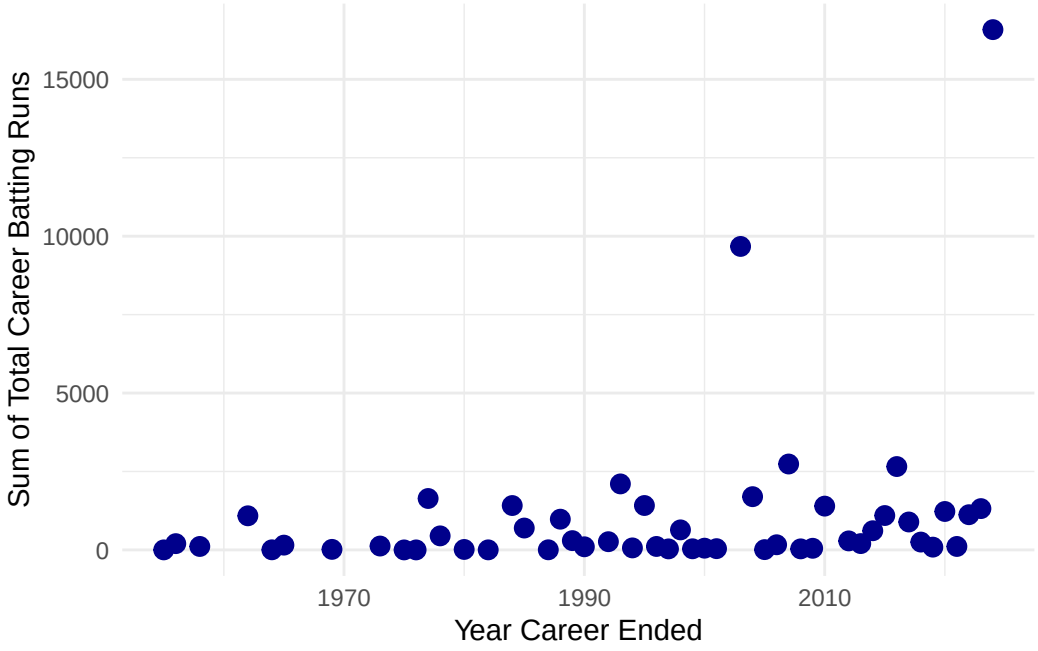


Figure 6: Total career batting runs for all players retiring in each year

player. The plot shows that most players retired with fewer than 1,000 career runs, and there is generally no strong positive trend across the data. However, there is a notable exception around 2015, where we see a slight uptick in the number of runs. This increase is short-lived, as the trend declines again shortly after, until 2024, when a large number of players retired, many of whom had accumulated over 1,000 career runs. This scatter plot provides useful information on individual player performances, which can be used to explore broader trends in the data.

Figure 6 shows a similar scatter plot, but in this case, each dot represents the total combined runs of all players retiring in a given year, with the y-axis now reflecting the sum of career runs for all retiring players in that year. Upon examining this graph, we observe some outliers, particularly in 2006 and 2024. Referring back to Figure 5, the 2006 outlier can be explained by a player retiring with around 6,800 runs. The 2024 outlier is due to the unusually high number of players retiring that year, including an exceptional player who ended their career with approximately 7,500 runs. Aside from these outliers, the overall trend remains relatively stagnant, with a slight increase again around 2015, though this quickly declines thereafter.

5 Discussion

5.1 Analysis Overview

In this paper, I employed Bayesian Logistic Regression to estimate the probability of Pakistan winning cricket matches, considering key factors such as the year of the match, the opponent, and the toss outcome. The model was specifically designed to understand how these variables influence Pakistan's chances of winning, while accounting for uncertainty and variability in the data. To ensure the analysis focused on relevant information, the data was cleaned and processed to include only Pakistan's matches, along with important player career statistics that could impact match outcomes. By incorporating these factors, the model provided a more complete understanding of Pakistan's performance dynamics.

The results of the analysis reveal useful intuitions into Pakistan's performance trends over time. For example, the strength of the opposing team played a significant role, with match outcomes showing a clear correlation between Pakistan's success and the competitive level of the opposition. Additionally, the toss outcome also emerged as influential factors, providing a deeper understanding of how certain conditions affect Pakistan's likelihood of winning. These findings underscore the importance of both internal (team performance) and external (opponent strength, match context) factors in shaping match outcomes, offering strategic information for future team preparation and performance analysis.

5.2 Effect of Opponent

One key finding from the analysis is the significant influence of the opposing team on Pakistan's match performance. As shown in Figure 2 and Table 1, the calculated coefficient values for each opposing team align with the patterns observed in Figure 3, where Pakistan has a higher success rate against teams with greater positive coefficient values, such as Bangladesh, and a lower success rate against teams with lower or negative coefficient values, such as South Africa. This correlation suggests that Pakistan's chances of winning are not only influenced by its own team performance but also by the strength of the opposition.

The model further emphasizes the variations in Pakistan's odds of winning depending on the opponent. For example, the coefficients for teams like Bangladesh, England, and Sri Lanka indicate more favorable matchups for Pakistan, suggesting that these teams have historically been more susceptible to Pakistan's strengths or less competitive overall. Conversely, the coefficients for teams like India and South Africa point to tougher challenges, with Pakistan's chances of winning being lower in these matchups. This pattern highlights the broader strategic landscape, where understanding the relative strength of the opponent becomes important for predicting outcomes. It underscores the need for teams to account not just for internal factors, such as player form or match conditions, but also for the competitive level of the opposition when preparing for matches. The results validate the intuition that the quality of the opposing team is a significant determinant of Pakistan's performance, and that adjusting tactics accordingly can improve chances against stronger teams. Ultimately, the analysis reinforces the importance of a strategic approach to match preparation, where the nature of the opposition is carefully considered alongside team strengths and other match variables.

5.3 Matches Won Over the years

When analyzing Figure 4 in relation to Figure 5 and Figure 6, the trends observed align with what we would expect based on the historical performance of Pakistan's cricket team. From around 2015, we see a noticeable increase in the number of games won by Pakistan. This can be explained by the presence of a number of veteran players who, with their vast experience, played a pivotal role in guiding the team to more victories. Around 2015, several seasoned players reached the peak of their careers or retired, bringing with them not only impressive career statistics but also leadership and expertise on the field. The presence of these experienced players undoubtedly helped stabilize the team's performance and contributed to the higher win rates during this period.

However, following 2015, there is a visible decline in Pakistan's performance, with a fluctuating number of wins per year. This drop in consistent victories is likely due to the retirement of key veteran players who had been essential to Pakistan's success. With their departure, a significant number of new, less experienced players entered the team. While these new players showed potential and skill, they lacked the depth of experience necessary to handle the pressures and complexities of international cricket. This transition phase, where fresh talent

was integrated into the team, led to inconsistency in match outcomes, which we observe as fluctuations in the win data post-2015.

Looking ahead to 2024, the retirement of another set of senior players further raises concerns about the future trajectory of Pakistan's performance. This continuous cycle of veteran players retiring and new players coming in without the same level of experience poses a challenge for the team. Although these new players bring in raw talent, the absence of experienced leaders and match-winners could contribute to a potential decline in overall performance. Given the historical patterns, we can reasonably anticipate a dip in the team's performance in the coming years unless the emerging players quickly adapt and gain the necessary experience to compete at the highest level. Therefore, while the retirement of key players is a natural part of the cycle, it highlights the ongoing challenge Pakistan faces in maintaining a competitive edge in international cricket.

5.4 Weaknesses and Limitations:

Despite the useful information gained, there are several limitations in the data and methodology that must be acknowledged. One significant challenge was the limited number of matches involving Pakistan that could be used to train the model. The data set included only 820 match observations for Pakistan, which, while informative, is relatively small when considering the complexity of cricket as a sport with numerous variables influencing match outcomes. This limited sample size could affect the robustness of the model's estimates, and in particular, the ability to generalize predictions across different conditions or time periods.

Additionally, there was a lack of detailed match statistics in the raw data, which constrained the scope of the analysis. For instance, while data on match outcomes (win/loss), toss results, and opponents were available, detailed statistics like the number of runs scored, wickets taken, or batting and bowling performances were only provided for the winning team. This lack of extensive match data—especially for losing teams—meant that I could not include key performance indicators like runs or wickets in the model, as many of these entries were missing or incomplete. These missing values could have influenced the accuracy of the model, and it would be prudent to consider these missing data points in future iterations of the analysis, potentially employing imputation methods or seeking alternative data sources.

Lastly, while the Bayesian Logistic Regression model is powerful, it assumes that the relationships between the predictors and the outcome are linear in nature. Although this is a reasonable assumption in many cases, cricket is a dynamic sport with complex interactions, and this model may not fully capture non-linear relationships or interactions between factors such as team composition, player injuries, or the impact of different match venues.

5.5 Next Steps

Despite these limitations, there are several directions for future work that could improve the analysis and expand our understanding of Pakistan's cricket performance.

- Improving the Data Set: One of the key areas for future research would be to gather a more extensive data set, particularly focusing on adding detailed match statistics, including runs, wickets, and player performances for both teams. This would allow for a more sophisticated model, potentially including these factors as predictors in a multivariate analysis. Additionally, expanding the sample size to include more matches across a broader time period could help provide more robust and generalizable results.
- Incorporating Player-Specific Metrics: Another avenue for improvement would be incorporating player-specific statistics into the model. By analyzing how the performance of individual players—batting and bowling averages, player form, and injury history—affects Pakistan's match outcomes, the model could be enhanced to offer a more granular understanding of what drives success or failure in cricket matches. Including player career data alongside match data would offer a deeper, more detailed look into the team's performance and could be used to identify potential strengths and weaknesses.
- Considering Interaction Effects: Future work could explore more sophisticated models that account for interactions between variables. For instance, the relationship between Pakistan's performance and the opponent could vary depending on the specific conditions of a given year, such as the emergence of new players, changes in team strategy, or alterations in the competitive landscape of international cricket. Exploring such interaction effects could provide a more refined view of Pakistan's performance over time.

Appendix

A Surveys, Sampling, and Observational Data

In this paper, we focus on a dataset of cricket matches involving Pakistan and their performance outcomes. However, a significant challenge in our analysis was the limited and often incomplete data available for each match, which made it difficult to fully answer the question of what affects Pakistan's overall match outcomes. Specifically, the dataset was missing detailed match statistics, such as the number of wickets per over, which player bowled in each over, individual player statistics, and detailed match events like runs and outs for both teams. This lack of granular data posed a substantial limitation to the analysis and highlighted the importance of accurate and complete data collection in sports analytics.

A.1 Data Collection

In an ideal scenario, the data would have been far more extensive, particularly with respect to the granular details of each match. To overcome the limitations of the current dataset, it would be essential to gather detailed match data through a combination of observational data and video analysis. This would involve either live or recorded footage of each match, from which key statistics could be manually or automatically extracted. Such an approach would include the following steps:

- `Detailed Over-by-Over Data`: Analyzing each over of the match would provide important data into the match's dynamics. This would involve recording key statistics, such as the number of runs scored, the number of wickets taken, the bowler and batter in each over, and any significant events, like boundary hits or wickets. This granular data would help in identifying the ebb and flow of the match and how different events influence Pakistan's chances of winning.
- `Player-Specific Statistics`: In addition to match-level statistics, it is essential to record individual player data. For every player involved, detailed statistics such as runs scored, wickets taken, the number of innings played, and the number of outs would be recorded. This information would allow for a more precise understanding of how the performance of individual players influences the match outcome. For instance, if certain players consistently score high or take wickets in key moments, this could provide useful information into the team's overall performance strategy.
- `Team-Specific Data`: Beyond individual player performance, it would be essential to capture team-level data for both the batting and bowling sides. This would include the number of runs scored by each team, the number of outs, and key tactical decisions like the batting order, field placements, and bowling changes. Understanding these aspects could allow for a better analysis of how team strategies affect the outcome of the match.

A.2 Annual Player Performance

Another important aspect that could improve the analysis is the collection of annual player performance data. While the current dataset allows for some intuition into match outcomes, it lacks detailed player statistics across multiple years. To address this, it would be helpful to build a dataset that combines match-level data for each player over the course of several years.

- `Annual Player Statistics`: For each player, aggregating data across multiple matches would provide a clearer picture of their performance over time. Key metrics such as the total number of runs scored, the number of wickets taken, the number of matches played, and the number of innings would be compiled. This would allow for a more refined analysis of individual player performance, considering factors like form and consistency.
- `Linking Player Performance to Match Outcomes`: By examining the link between individual player statistics and match outcomes, it would be possible to identify trends that might not be evident at the match level. For example, players with higher annual runs might be correlated with higher chances of winning for the team, or certain bowlers who take more wickets in specific conditions could become key to success. This approach would allow for more precise modeling of match outcomes based on both individual and team performance.
- `Player and Opponent Comparisons`: To better understand how player performance interacts with match outcomes, it would be useful to collect data on player performances not just for Pakistan, but also for opposing teams. By comparing the performances of Pakistan's players with those of the opposing team's players, we could gain knowledge on how the relative strength of players affects the result. For example, if Pakistan's key players perform well but the opposing team's star players outperform them, it could explain a loss despite good individual performances.

A.3 Simulation Study

A key aspect of ensuring the robustness of the data is to evaluate the impact of potential sampling bias on performance predictions. By running a simulation study, we could artificially create datasets with varying levels of completeness and bias (e.g., underrepresentation of certain match conditions or players) and test how these biases affect the ability to predict match outcomes. For instance, simulations could explore scenarios where certain conditions, such as playing at home or under specific weather conditions, are underrepresented in the data. By comparing the outcomes of these simulations with real-world performance predictions, we can better understand how bias in sampling influences model accuracy and the generalizability of results.

A.4 Bias-Adjustment Techniques

In addition to simulation, statistical techniques such as propensity score matching or inverse probability weighting could be applied to adjust for biases in the observed data. These methods would help to correct for any over- or under-representation of certain conditions and ensure that the analysis is reflective of a broader set of match scenarios. For example, if the dataset is skewed towards games played against stronger teams, these techniques could help balance the data to ensure that the analysis is not overly influenced by these high-stakes matchups.

B Model details

B.1 Model assumptions

- `Independence of Observations`: We assume that the results of each match are independent. This assumption may be violated in cases where teams play multiple matches within a short period, but the model does not account for such temporal dependencies.
- `Logistic Relationship`: The model assumes a linear relationship between the predictors (year, opponent, toss win) and the log-odds of Pakistan winning. This assumption may not capture more complex non-linear relationships, which would require more sophisticated models, such as generalized additive models (GAMs).
- `No Interaction Effects`: The model currently does not include interaction terms (e.g., how the effect of winning the toss might differ by opponent). Interaction effects could be considered if they were hypothesized to be important, but they are omitted here for simplicity.
- `Reference Category for team2`: The model assumes that the effect of playing against a reference opponent (the one excluded from the dummy variables for team2) is the baseline against which the other teams are compared. The choice of reference category is important and could influence the interpretation of the coefficients.

B.2 Model limitations

- `Omitted Variables`: The model does not account for other potentially important variables, such as player injuries, weather conditions, or venue specifics. These factors could influence the match outcome but are not included in the current model.
- `Data Representation`: The model is based on matches involving Pakistan, so it does not generalize to other teams or international cricket broadly. Additionally, the dataset may have biases (e.g., a higher number of matches against certain opponents), which could affect the accuracy of the predictions.

- `Data Sparsity`: The dataset available for analysis may not be large enough to fully capture all the complexities of cricket matches, especially with respect to less frequent events or rare outcomes. This could lead to uncertainty in the model estimates, particularly for smaller categories or outcomes that appear less frequently.
- `Data Granularity`: The model uses year as a continuous variable. However, time trends in cricket might be better captured with more granular temporal variables (e.g., by series or tournament), as the year-by-year analysis may mask shorter-term trends.

B.3 Model validation

We split the data into training and test sets to evaluate the predictive performance of the models. The first model was trained on a subset of the data, and its predictions were tested on the remaining matches in the test set. Specifically, we used out-of-sample testing, where the model was trained on the training set and evaluated on the test set to assess its ability to generalize to new, unseen data.

For the first model, we predicted the probability of a match outcome (e.g., whether Pakistan would win) for each match in the test set. Using a threshold of 0.5, we classified these probabilities into binary outcomes (1 for a predicted win, 0 for a predicted loss). To evaluate model performance, we constructed a confusion matrix comparing the predicted classes to the actual outcomes of the test set. From this confusion matrix, we calculated accuracy as the proportion of correct predictions (i.e., the sum of true positives and true negatives divided by the total number of predictions). This provided an overall measure of how well the model performed in predicting match outcomes. The result of these tests is summarized in Table 2 in the form of a confusion matrix and accuracy.

We repeated the same procedure for the second model, predicting the match outcome based on a different set of features (e.g., year and opponent). The accuracy for both models was computed and reported separately.

This approach provided a robust evaluation of each model's predictive performance on unseen data, with the confusion matrix and accuracy offering information on the model's classification capabilities. While accuracy provides a good indication of overall model performance, we also recognize that other metrics such as precision, recall, and F1 score might be more informative in the case of imbalanced classes.

Table 2: Confusion matrix and Accuracy measurements for model

```
      Predicted
Actual  0  1
     0 18  4
     1  5  3


Model Accuracy: 0.70
```

B.4 Model diagnostics

Fit Diagnostics:

The mean posterior predictive distribution (mean_PPD) shown in Table 3 is the mean of the posterior distribution of the predicted probabilities for Pakistan winning a match. The value of 0.4 indicates that, on average, the model predicts Pakistan has a 40% chance of winning a match based on the predictors.

MCMC Diagnostics:

Table 4 shows diagnostic measurements of the sampling process to assess whether the model has converged and whether the samples are reliable.

- `mcse (Monte Carlo Standard Error)`: The MCSE indicates the standard error of the Monte Carlo estimates for each parameter. Smaller values are better, as they show that the estimates are stable and not too variable. Most of the parameters here have small MCSE values (close to 0), indicating stable estimates.
- `Rhat`: Rhat is a measure of convergence. An Rhat value of 1.0 indicates perfect convergence, meaning the chains for the MCMC sampling are mixing well and that the estimates for each parameter are reliable. All Rhat values here are exactly 1.0, which suggests that the sampling chains converged properly.
- `n_eff (Effective Sample Size)`: This is a measure of how many independent samples are available for each parameter. Larger values indicate better sampling. Here, most parameters have high effective sample sizes (ranging from ~1600 to ~3000), suggesting that the model has gathered a sufficiently large and diverse set of samples for each parameter, ensuring robust inference.

Table 3: Model posterior predictive distribution

	mean	sd	10%	50%	90%
mean_PPD	0.38	0.06	0.31	0.38	0.45

Table 4: Model MCMC statisics

	mcse	Rhat	n_eff
(Intercept)	1.46	1	3121
year	0.00	1	3135
team2Bangladesh	0.03	1	1965
team2England	0.02	1	1670
team2India	0.03	1	2031
team2Ireland	0.41	1	1648
team2New Zealand	0.02	1	1887
team2South Africa	0.02	1	1853
team2Sri Lanka	0.02	1	1539
team2West Indies	0.02	1	1809
team2Zimbabwe	0.03	1	2715
toss_win	0.01	1	3283

References

Arel-Bundock, Vincent. 2022. "modelsummary: Data and Model Summaries in R." *Journal of Statistical Software* 103 (1): 1–23. https://doi.org/10.18637/jss.v103.i01.

Cricsheet. 2024. "Cricsheet: Cricket Data for Everyone." https://cricsheet.org.

ESPN Cricinfo. 2024. "ESPN Cricinfo." https://www.espncricinfo.com.

Hyndman, Rob, Charles Gray, Sayani Gupta, Timothy Hyndman, Hassan Rafique, and Jacquie Tran. 2024. *Cricketdata: International Cricket Data*. https://github.com/robjhyndman/cricketdata.

Kuhn, and Max. 2008. "Building Predictive Models in r Using the Caret Package." *Journal of Statistical Software* 28 (5): 1–26. https://doi.org/10.18637/jss.v028.i05.

Müller, Kirill, and Hadley Wickham. 2023. *Tibble: Simple Data Frames*. https://CRAN.R-project.org/package=tibble.

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, Ian Cook, Nic Crane, Dewey Dunnington, Romain François, Jonathan Keane, Dragoș Moldovan-Grünfeld, Jeroen Ooms, Jacob Wujciak-Jens, and Apache Arrow. 2024. *Arrow: Integration to 'Apache' 'Arrow'*. https://CRAN.R-project.org/package=arrow.

Vehtari, Aki, Daniel Simpson, Michael Betancourt, Paul B. Gelman, Jonathan Goodrich, Ben McElreath, Jonah Gabry, and Peter J. L. Chodera. 2022. *Rstanarm: Bayesian Applied*

Regression Modeling via Stan. https://mc-stan.org/rstanarm/.

Wickham, Hadley. 2011. "Testthat: Get Started with Testing." *The R Journal* 3: 5–10. https://journal.r-project.org/archive/2011-1/RJournal_2011-1_Wickham.pdf.

———. 2016. *Ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York. https://ggplot2.tidyverse.org.

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, Kirill Müller, and Davis Vaughan. 2023. *Dplyr: A Grammar of Data Manipulation*. https://CRAN.R-project.org/package=dplyr.

Wickham, Hadley, Jim Hester, and Jennifer Bryan. 2024. *Readr: Read Rectangular Text Data*. https://CRAN.R-project.org/package=readr.

Wickham, Hadley, Davis Vaughan, and Maximilian Girlich. 2024. *Tidyr: Tidy Messy Data*. https://CRAN.R-project.org/package=tidyr.

Xie, Yihui. 2023. *Knitr: A General-Purpose Package for Dynamic Report Generation in r*. https://cran.r-project.org/package=knitr.

Zhu, Hao. 2024. *kableExtra: Construct Complex Table with 'Kable' and Pipe Syntax*. https://CRAN.R-project.org/package=kableExtra.