Analyzing the Pakistan Cricket Team's Performance (2002-2024)*

My subtitle if needed

Muhammad Abdullah Motasim

December 3, 2024

First sentence. Second sentence. Third sentence. Fourth sentence.

Table of contents

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

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 insights into their journey and areas for strategic improvement.

Despite an extensive focus on player statistics and match highlights, there remains a lack of systematic statistical analysis to evaluate long-term performance trends. This paper fills this gap by analyzing data from 2002 to 2024 to understand what factors contribute to Pakistan's success in international cricket.

We employ a linear regression model to analyze match outcomes, considering variables like year, opponent, and toss outcomes. The focus is on identifying trends and predicting the probability of victory under varying conditions. Our findings not only contribute to the literature on sports analytics but also provide actionable insights for coaches, analysts, and cricket administrators not only in Pakistan, but around the world.

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 Linear 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), tinytable (Arel-Bundock 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 comprehensive 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.

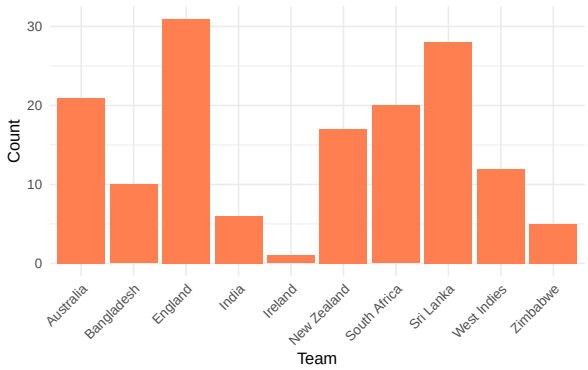
2.4 Variable Analysis

Figures Figure 1 displays the count of several important match statistics. From these, we can observe that Pakistan faces a variety of opponents, with the most frequent ones being England, Sri Lanka, and Australia. Notably, Pakistan has had a challenging record against Australia, losing a significant number of matches, while many of their matches ended in draws.

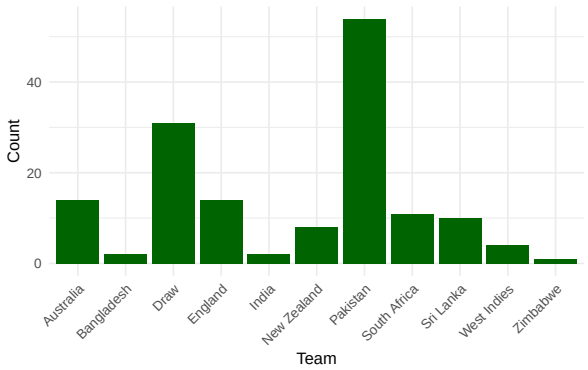
Regarding coin toss outcomes, Sri Lanka and South Africa appear to have won a substantial number of tosses. However, it's important to note that while the graphs for match outcomes and tosses won may give the impression that Pakistan was victorious in most of these events, we must keep in mind that the 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. These figures represent Pakistan's performance specifically, not that of other teams.

The following two graphs show the number of matches and tosses that Pakistan won or lost. The win-loss ratio is approximately 0.5, indicating 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.

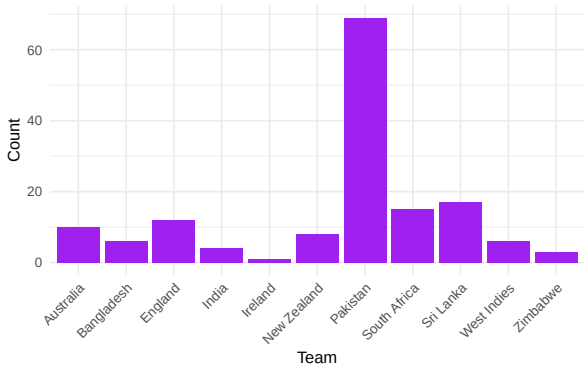
Finally, the last graph illustrates the number of matches Pakistan played each year. We can see a peak in 2016, with a gradual decline in matches played over the subsequent years, culminating in a sharp drop in the most recent year—comparable to the period during the COVID-19 pandemic.



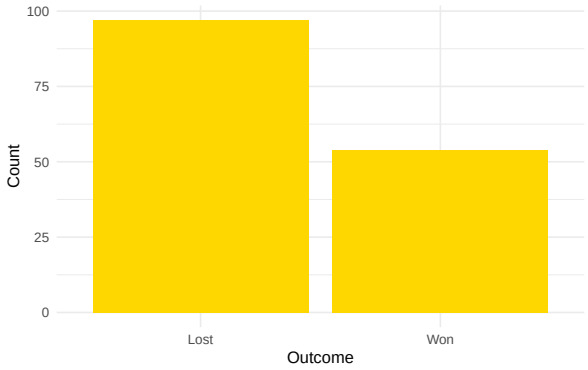
(a) Pakistan Opponent Team



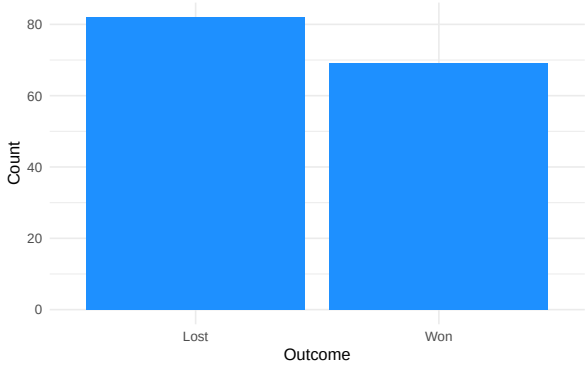
(b) Matches Won by Each Team



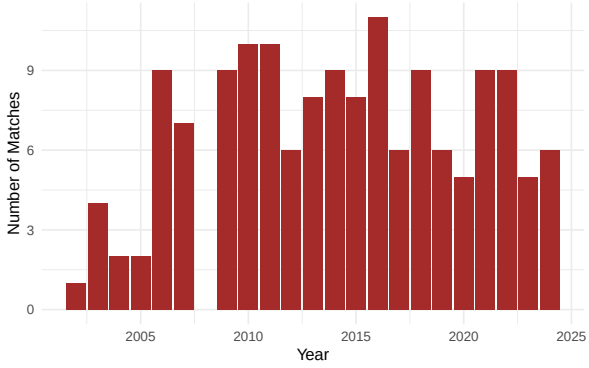
(c) Tosses Won by each Team



(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

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 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 insights 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:

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

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

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 of the model coefficients is assigned a Normal prior with a mean of 0 and a standard deviation of 2.5. This reflects the assumption that we expect the effects of the predictors to be centered around zero, with a reasonable amount of uncertainty (a wide spread due to the scale of 2.5) about their true values.
- `Prior for the Intercept`: A normal prior with the same parameters as the coefficients (mean = 0, scale = 2.5) is used for the intercept. The intercept represents the baseline log-odds of Pakistan winning a match when all other predictors are zero.
- `Prior for the Auxiliary Parameter (prior_aux)`: This parameter represents the dispersion of the logistic regression model. An Exponential prior with a rate of 1 is used, which is typical in Bayesian models for logistic regression. The Exponential prior allows for a distribution with only non-negative values, reflecting uncertainty in the variance of match outcomes, with higher values suggesting more variability in outcomes.

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, offers 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 presents the predicted values of the coefficients. This 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.

Analyzing some of these coefficients further, The results from the regression model are as follows:

- **Intercept**: The intercept of -55.94 (with a standard error of 80.63) represents the log-odds of Pakistan winning a match when all other variables are set to their reference categories. This value is quite large in magnitude, but it is not directly interpretable in isolation. The negative value suggests a very low baseline probability of winning in the absence of other influencing factors.
- **Year**: The coefficient for year is 0.03 (standard error = 0.04). This positive coefficient suggests that, as time progresses, the probability of Pakistan winning a match increases very slightly. However, the small magnitude of the coefficient indicates that this effect is minimal. The relatively narrow standard error suggests that the estimate is reasonably precise, but the practical significance of this trend is modest.

Table 1: Coefficient estimates and standard errors

	Model
(Intercept)	−55.94
	(80.63)
year	0.03
	(0.04)
team2Bangladesh	3.62
	(1.31)
team2England	1.18
	(0.72)
team2India	0.24
	(1.44)
team2Ireland	19.77
	(15.67)
team2New Zealand	0.74
	(0.86)
team2South Africa	0.29
	(0.90)
team2Sri Lanka	1.09
	(0.72)
team2West Indies	2.26
	(0.86)
team2Zimbabwe	2.58
	(1.48)
toss_win	0.60
	(0.45)

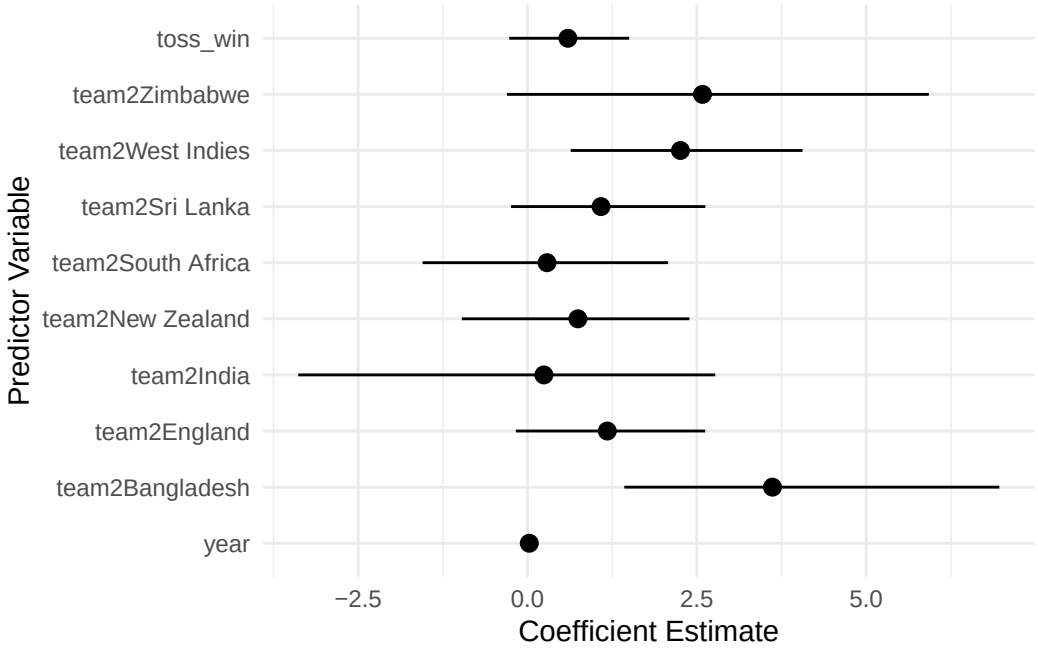


Figure 2: Coefficient estimates and 95% confidence intervals

- **Opponent**: The coefficients for the opponent variable represent how the likelihood of Pakistan winning a match changes when facing different teams. These coefficients should be interpreted relative to the reference team, which is not explicitly listed here but is assumed to be the team with the lowest or most neutral effect:
  - `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 showcases the number of wins, losses and draws Pakistan had against each opponent from the data set.

5 Discussion

5.1 First discussion point

If my paper were 10 pages, then should be be at least 2.5 pages. The discussion is a chance to show off what you know and what you learnt from all this.

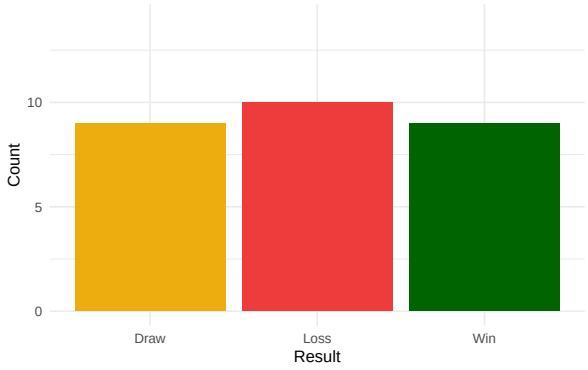
5.2 Second discussion point

Please don't use these as sub-heading labels - change them to be what your point actually is.

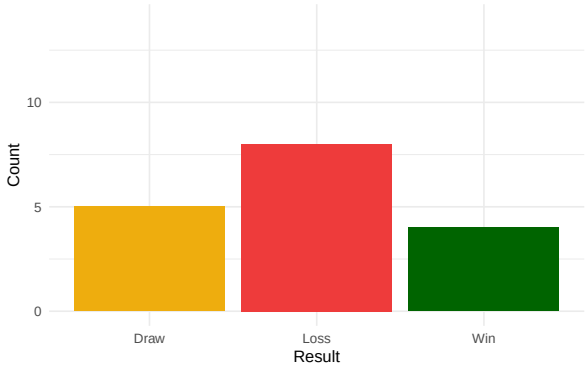
5.3 Third discussion point

5.4 Weaknesses and next steps

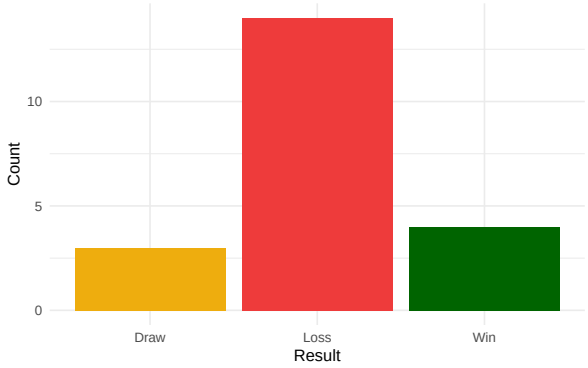
Weaknesses and next steps should also be included.



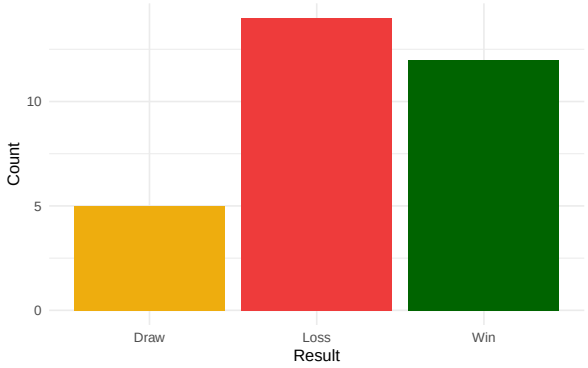
(a) Sri Lanka



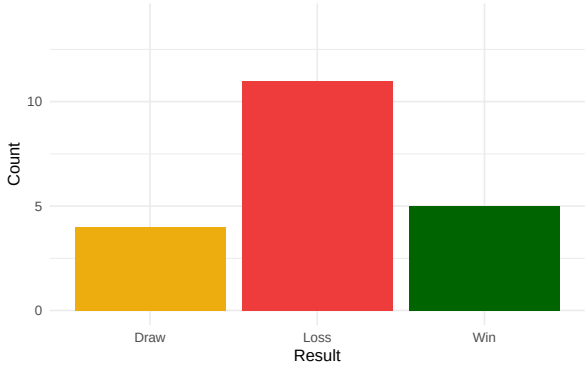
(b) New Zealand



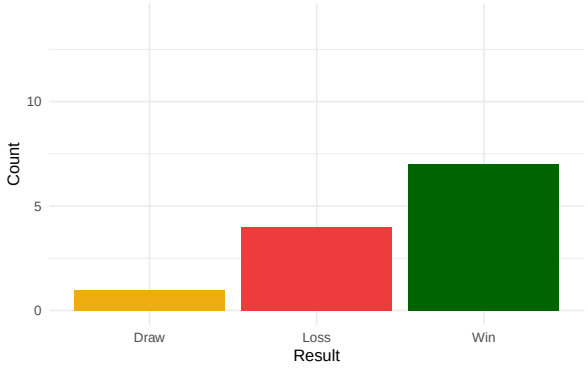
(c) Australia



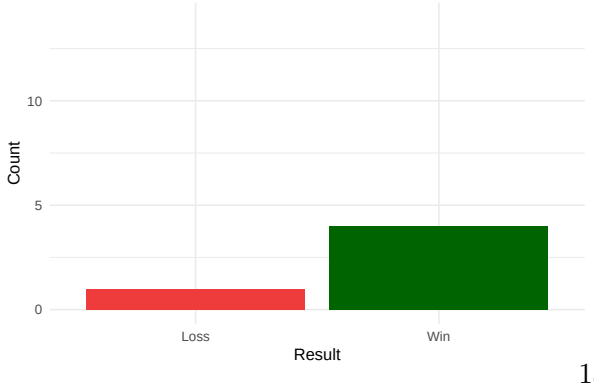
(d) England



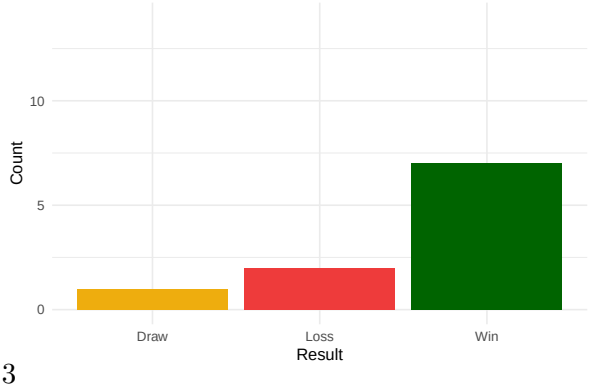
(e) South Africa



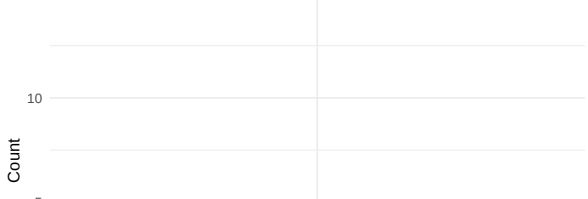
(f) West Indies

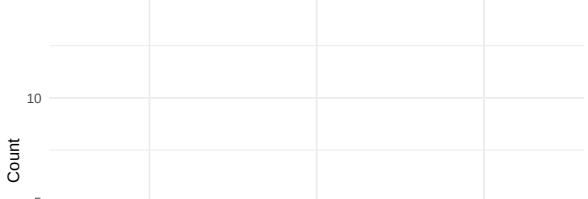


(g) Zimbabwe



(h) Bangladesh





Appendix

A Additional data details

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 insights into 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
```

Table 3: Model posterior predictive distribution

	mean	sd	10%	50%	90%
mean_PPD	0.38	0.05	0.31	0.38	0.45

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 4: Model MCMC statisics

	mcse	Rhat	n_eff
(Intercept)	1.51	1	2930
year	0.00	1	2943
team2Bangladesh	0.03	1	2012
team2England	0.02	1	1660
team2India	0.04	1	1964
team2Ireland	0.36	1	1882
team2New Zealand	0.02	1	2058
team2South Africa	0.02	1	1936
team2Sri Lanka	0.02	1	1629
team2West Indies	0.02	1	1835
team2Zimbabwe	0.03	1	2575
toss_win	0.01	1	2990

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.

———. 2024. *Tinytable: Simple and Configurable Tables in 'HTML', 'LaTeX', 'Markdown', 'Word', 'PNG', 'PDF', and 'Typst' Formats.* https://CRAN.R-project.org/package=tinytable.

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.