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

My subtitle if needed

Muhammad Abdullah Motasim

December 2, 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), patchwork (Pedersen 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 and Figure 2 display 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.

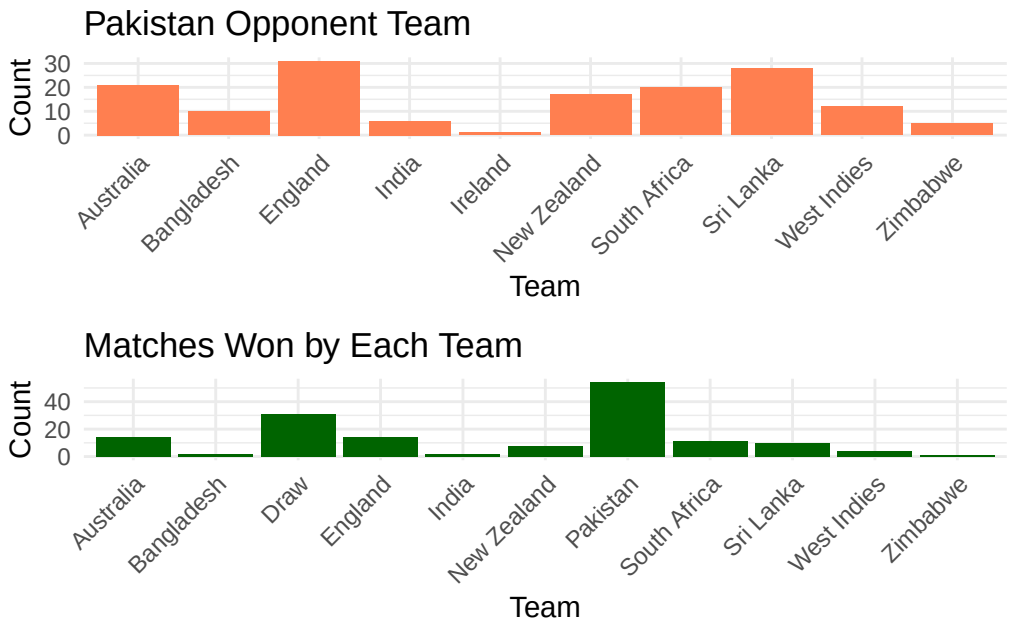

Figure 1: Counts for match variables of interst

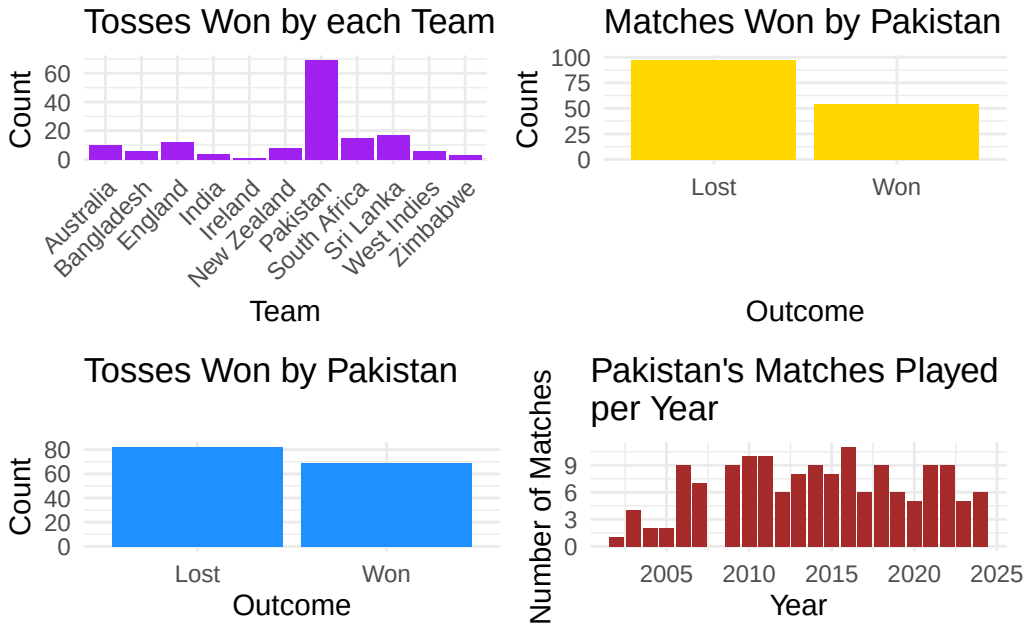

Figure 2: Counts for match variables of interst

3 Model

The goal of our modelling strategy is twofold. Firstly,…

Here we briefly describe the Bayesian analysis model used to investigate… Background details and diagnostics are included in Appendix B.

3.1 Model set-up

Define y_i as the number of seconds that the plane remained aloft. Then β_i is the wing width and γ_i is the wing length, both measured in millimeters.

$$y_i|\mu_i, \sigma \sim \text{Normal}(\mu_i, \sigma) \tag{1}$$
$$\mu_i = \alpha + \beta_i + \gamma_i \tag{2}$$
$$\alpha \sim \text{Normal}(0, 2.5) \tag{3}$$
$$\beta \sim \text{Normal}(0, 2.5) \tag{4}$$
$$\gamma \sim \text{Normal}(0, 2.5) \tag{5}$$
$$\sigma \sim \text{Exponential}(1) \tag{6}$$

We run the model in R (R Core Team 2023) using the `rstanarm` package of Vehtari et al. (2022). We use the default priors from `rstanarm`.

3.1.1 Model justification

We expect a positive relationship between the size of the wings and time spent aloft. In particular…

We can use maths by including latex between dollar signs, for instance θ.

4 Results

Our results are summarized in Table **??**.

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.

Appendix

A Additional data details

B Model details

B.1 Posterior predictive check

In **?@fig-ppcheckandposteriorvsprior-1** we implement a posterior predictive check. This shows…

In **?@fig-ppcheckandposteriorvsprior-2** we compare the posterior with the prior. This shows…

Examining how the model fits, and is affected
by, the data

B.2 Diagnostics

?@fig-stanareyouokay-1 is a trace plot. It shows… This suggests…

?@fig-stanareyouokay-2 is a Rhat plot. It shows… This suggests…

Checking the convergence of the MCMC algo-
rithm

References

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.

Pedersen, Thomas Lin. 2024. *Patchwork: The Composer of Plots*. https://CRAN.R-project.org/package=patchwork.

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.