Predicting the winner of IPL based on league standing table

Aditya Ranade

December 15, 2021

1 Introduction

Sports is one of the most popular form of entertainment. Various sporting leagues have come up which are played throughout the world. We will consider one such sporting league for the project. Indian Premier League (IPL) is an annual T20 (20 overs each side) cricket league played in India which was started in 2008. For most of the years till 2021, the league was played between eight teams based out of various cities in India in a double round robin format except for a couple of years when it was played with 10 teams.

The league stage consists of 14 games for each team and the top 4 teams at the end of league stage play the play-offs to determine the champions. For first three seasons (2008-2010), the league followed a traditional semi finals and final approach after the league stages. In order to reward the top two teams of league stages for being consistent through the league stage, play-offs were redesigned from 2011 to give the top two teams of league stage, two chances to make the finals. Figure 1 in figures section shows the the playoff structure. Due to the new design of the playoffs, teams placed 1^{st} or 2^{nd} at the end of league stage cannot finish below position 3 at the end of the tournament whereas teams placed 3^{rd} or 4^{th} at end of league stage can finish at any position from 1 to 4 at the end of the tournament.

In this project, we are interested to determine the winner of the tournament based on commonly available information through points table after the league stage for the top four teams.

2 Data

The data considered for this project is available from the league standing table which is commonly referred to as points table (at the end of the league stage). The data is obtained through IPL website (iplt20.com/points-table) for all the years 2010 upto 2021. Please note data for years 2008 upto 2010 is available but not considered since the league followed the traditional semi finals and final match approach in those years.

The variables considered are as follows

- **year**: Indicates the year from which the observation is
- **result**: Eventual position of the team after the end of tournament. Possible values {1,2,3,4}.

- **champion**: Binary variable derived from result variable which takes value 0 if the team does not win the league and 1 if the team wins the league. Only one of the four observations for any given year can have value 1.
- **seed**: Team standing at the end of league stage which can be treated as 'seed' which is ordinal variable. Possible values {1,2,3,4} with no two teams from a year can have same number.
- **nrr**: Net run rate at the end of league stage. $nrr = \left(\frac{\text{Runs scored}}{\text{overs played}}\right) - \left(\frac{\text{Runs conceded}}{\text{overs bowled}}\right)$ Can take any number on real line but usually between -3 and 3.
- **last5**: Number of matches won out of the last 5 league matches. Possible values are {0,1,2,3,4,5}. This variable can be considered as a proxy for 'form' as higher value indicates the team is in form coming into the playoffs.

Data from 2011 to 2020 is used for the estimation of parameters and data from 2021 is used for prediction. A sample of the data for year 2017 can be seen in Table 2

2.1 Exploratory data analysis

From the structure of the playoffs (Figure 1), it seems the team finishing in the top 2 positions of the league stage (seed 1 and 2) has an advantage in winning the league (since they get 2 chances to be in the finals).

The result variable indicates the position where the team finished after the end of tournament. Hence result variable can be considered to have a Categorical distribution

$$\text{result}_{\text{seed}} \sim \begin{cases} \text{Categorical}(3, \underset{\sim}{\pi}), & \text{seed} = 1, 2 \\ \text{Categorical}(4, \underset{\sim}{\pi}), & \text{seed} = 3, 4 \end{cases}$$

where $\underset{\sim}{\pi}$ is the vector of probabilities of finishing at each position.

Team seeded first and second can finish the tournament at any position from 1 to 3 whereas team seeded third and fourth can finish the tournament at any position from 1 to 4 hence the two different set of parameters of the Categorical distribution depending on the value of seed variable.

Considering uninformative Dirichlet prior gives us a simple independent Categorical - Dirichlet model for the distribution of the positions the team seeded i where $i = 1, \ldots, 4$ can finish and the corresponding probabilities of finishing at each of the possible position. The credible intervals for the probabilities of finishing the league at positions 1 to 4 can be found in Figure 2. Teams seeded first and second cannot finish below position 3 hence no lines for position 4 (probability of finishing fourth at the conclusion of league) for the teams seeded first and second.

The histogram for net run rate variable (nrr) and number of last 5 league matches won (last5) variables can be seen in Figure 3. Both appear symmetric and do not indicate any concerns.

3 Methods

3.1 Model Rationale

A Multinomial logistic regression to predict the end position of the team qualified for the playoffs (seed 1 through 4) would be ideal. However the assumption of this model would be that each team who qualified for playoffs (seed) has a non zero probability of finishing the tournament at all the positions from 1 to 4. However, due to the playoff structure, the team seeded 1 or 2 cannot finish the tournament at position below 3. Hence the Multinomial logistic model is ruled out.

The variable champion is a binary variable derived from result variable which indicates if the team finished tournament as champion.

$$\text{champion} = \begin{cases} 1, & \text{result} = 1 \\ 0, & \text{result} = 2, 3, 4 \end{cases}$$

All the teams who qualified for the playoffs have a chance to win the tournament so champion variable can take value 0 or 1 for each of the team. A Bayesian approach to the logistic regression on the champion variable will be a valid model.

3.2 Model

For the logistic regression, the response variable (Y) will be the binary variable champion whereas seed, nrr and last 5 variables will be the explanatory variables. Seed variable is a ordinal variable with 4 categories and is used in the analysis through dummy variables. The dummy variable are seed2 (seed = 2), seed3 (seed = 3) and seed4 (seed = 4) whereas the base category will be seed = 1 which gets absorbed in the intercept.

$$Y_i|\mathbf{X}, \underset{\sim}{\beta} \sim bernoulli(\pi_i) \ , \ i = 1, 2, ..., n$$

$$\pi_i = \frac{exp(\beta_0 + \beta_1 * seed2 + \beta_2 * seed3 + \beta_3 * seed4 + \beta_4 * nrr + \beta_5 * last5)}{1 + exp(\beta_0 + \beta_1 * seed2 + \beta_2 * seed3 + \beta_3 * seed4 + \beta_4 * nrr + \beta_5 * last5)}$$

$\mathbf{X}$ is the matrix containing the values of seed, nrr and last5 variables.

Uninformative priors are considered for the parameters as follows

$$\beta_0 = \beta_{intercept} \sim N(0, 5)$$
$$\beta_1 = \beta_{seed2} \sim N(0, 5)$$
$$\beta_2 = \beta_{seed3} \sim N(0, 5)$$
$$\beta_3 = \beta_{seed4} \sim N(0, 5)$$
$$\beta_4 = \beta_{nrr} \sim N(0, 5)$$
$$\beta_5 = \beta_{last5} \sim N(0, 5)$$

The priors were chosen with center 0 which indicates the corresponding variable does not impact the chances of winning the championship. After observing the data, if the variable has an effect on the team's chances of winning the league, posterior distribution will reflect it accordingly in the credible intervals.

3.3 Model Fitting

The model was run using the brms package in R version 4.1.2 . A 4 chain MCMC was run with 4000 iterations for each chain with a burn-in period of 1000 iterations. This yields 12,000 MCMC samples for each of the parameters $\beta_{intercept}, \beta_{seed2}, \beta_{seed3}, \beta_{seed4}, \beta_{nrr}, \beta_{last5}$. The potential scale reduction factor for the parameters which can be found in Table 2 are all 1. In addition to scale reduction factor, the trace plots as seen in Figure 4 indicates well mixed chains. The potential scale reduction factor and the trace plots do not indicate a lack of convergence of the Markov chains.

4 Results

4.1 Parameter estimates

The posterior distributions of the unknown parameters along with their estimates were obtained using MCMC. The posterior distribution for the model parameters $\beta_{intercept}, \beta_{seed2}, \beta_{seed3}, \beta_{seed4}, \beta_{nrr}$, β_{last5} can be seen in Figure 5 . The parameter estimates can be found in Table 3 which are in logit scale (log odds). The effect of each variable on the team's chances of winning the tournament are best understood on the probability scale. The median and 95% credible interval of the parameter estimates converted from the logit scale to the probability scale can be found in Table 4

4.2 Effects of parameters

On the logit scale, the posterior 95% credible intervals for all the parameters except β_{seed4} includes 0 which implies that on probability scale, 0.5 is included in the credible intervals. This indicates except seed4 variable, the other variables do not significantly impact the chances of team winning the league. However, if we take a carefully look at the intervals, 0 (on logit scale) is close to either lower or the upper bound of the interval. Hence we will interpret the median of the intervals.

The effect of the seed variable is as follows

- Team seeded second has around 35% higher chance of winning the championship compared to being seeded first when other variables (nrr and last5) are held constant.
- Team seeded third has around 30% lower chance of winning the championship compared to being seeded first when other variables (nrr and last5) are held constant.
- Team seeded fourth has around 49.91% lower chance of winning the championship compared to being seeded first when other variables (nrr and last5) are held constant.

For the nrr variable, every unit increase in the net run rate for the team coming into the playoffs, the chances of winning the league increases by 29% when other variables (seed and last5) are held constant.

For the last5 variable, every addition win (out of the last 5 league games) that the team comes to the playoffs with, increases the chances of winning the league by 13% when other variables (seed and nrr) are held constant.

4.3 Conditional Effects

Figure 6 gives the conditional effects for the seed variable. The plot gives the probability intervals for teams seeded 1 to 4 winning the tournament conditioned on the average values of nrr and last5 variable. This plot reiterates the information we get from the parameter estimates. The highest chance of winning the tournament is to the team seeded second followed by team seeded first, third and fourth respectively.

4.4 Prediction

The model was estimated using data from year 2011 to 2020. We will try to predict the winner for year 2021 based on this model. Table 5 gives us information from league standing table, the probability of team winning the tournament. Since the model gives independent probabilities for each team, they do not follow the law of total probability. We achieve this by dividing the probability of each team winning by the sum of probabilities of each of the four teams winning. This results in team seeded 2 being assigned highest probability of winning the tournament and it turns out to be a correct prediction. However, the effect of seed on the prediction is such that the model will probably always predict the team seeded 2 as the team that will win the tournament.

5 Discussion

The model considered in this project has a drawback that it considers each observation as independent. In reality, the four observations in any given year are dependent on each other whereas the observations between the years are independent. We can bypass this flaw by dividing the probabilities of each observation in a year by the sum of probabilities of the four observations in the same year. It would be worthwhile to build this dependence structure within the model.

The posterior credible intervals of all the parameters except the parameter corresponding to seed4 contains 0 on the logit scale which indicates there is no significant effect of those variables (namely seed2, seed3, nrr and last5) on the chances of winning the tournament. This can be due to limited data. In the 10 years, team seeded second has gone on to win the tournament six times hence the model will probably always predict the team seeded second to win the league. It would be interesting to run this model after having observations from few more years.

The R code used for this analysis is attached in the Appendix

6 Improvements and Future Work

A better model would be to simulate the playoffs after considering the data from the tournaments about the matches between each of the teams and try to determine team strength which would help with calculating the probabilities of teams winning against each other.

Another improvement could be to consider the interaction of teams with the seeds. However IPL rules are set in such a way that the entire teams can change typically every 3 years as the league follows auction process to select players for the whole squad every 3 years. This probably results in huge variation among the data.

7 Figures

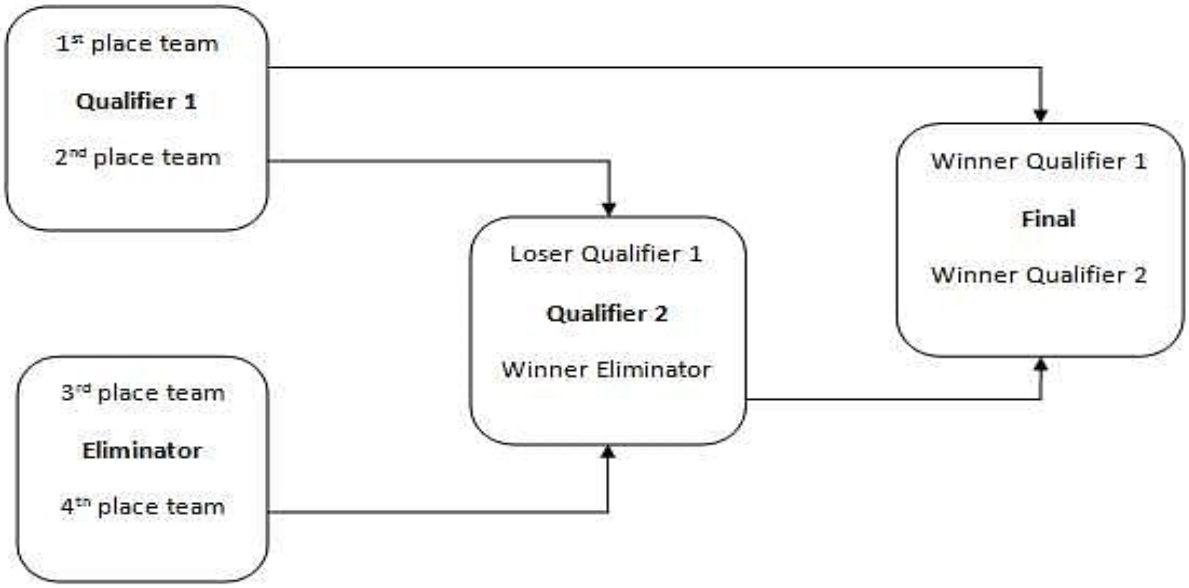

Figure 1: Playoff structure

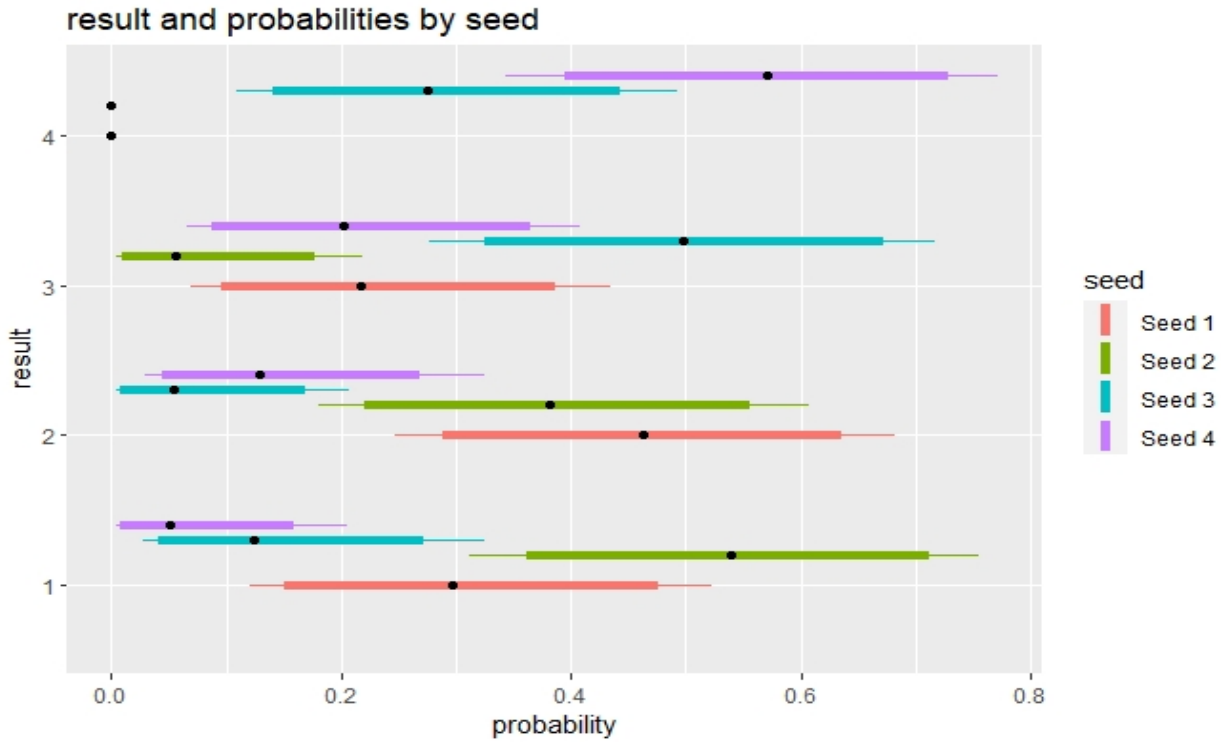

Figure 2: End position probabilities by seed

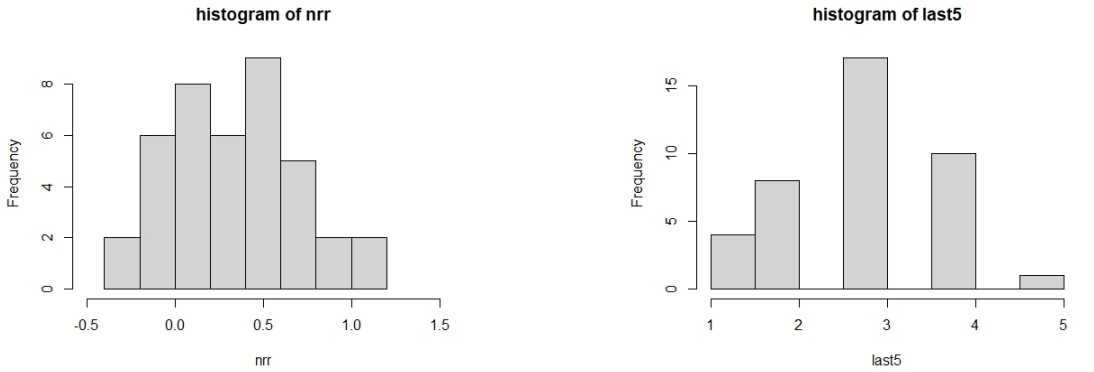

Figure 3: histogram of nrr and last5

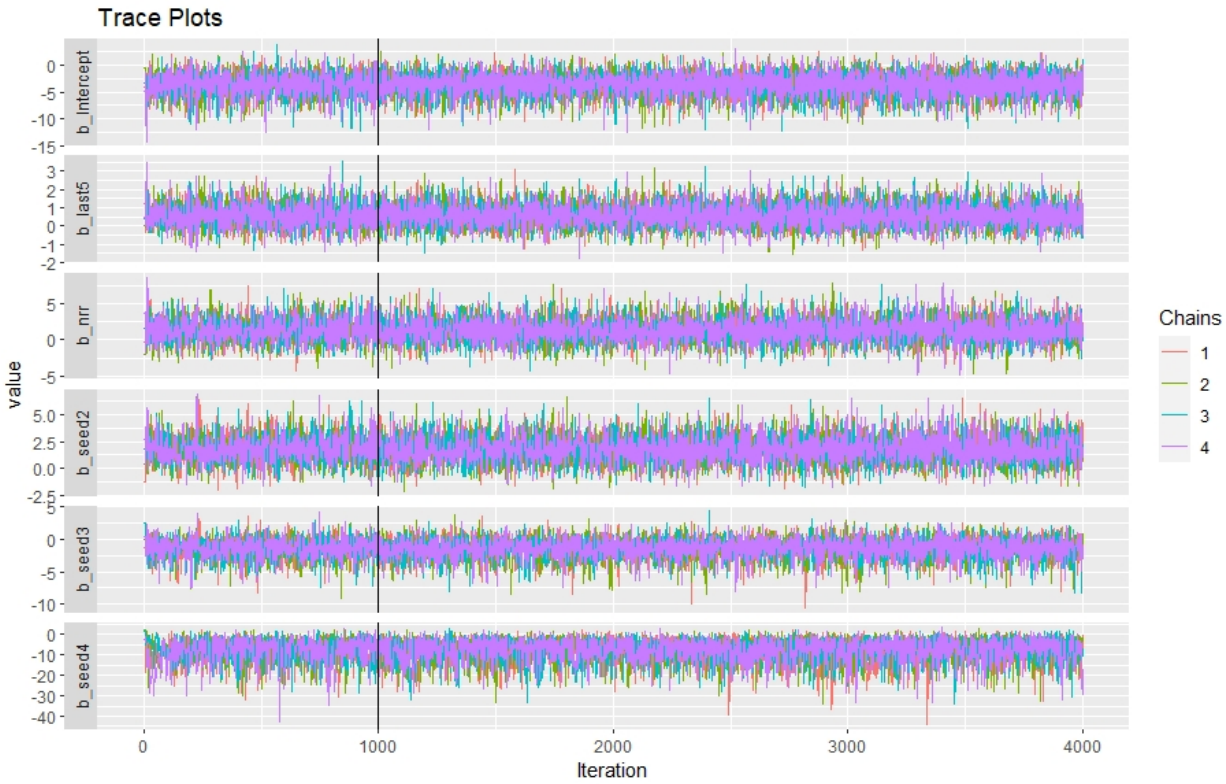

Figure 4: Trace plots

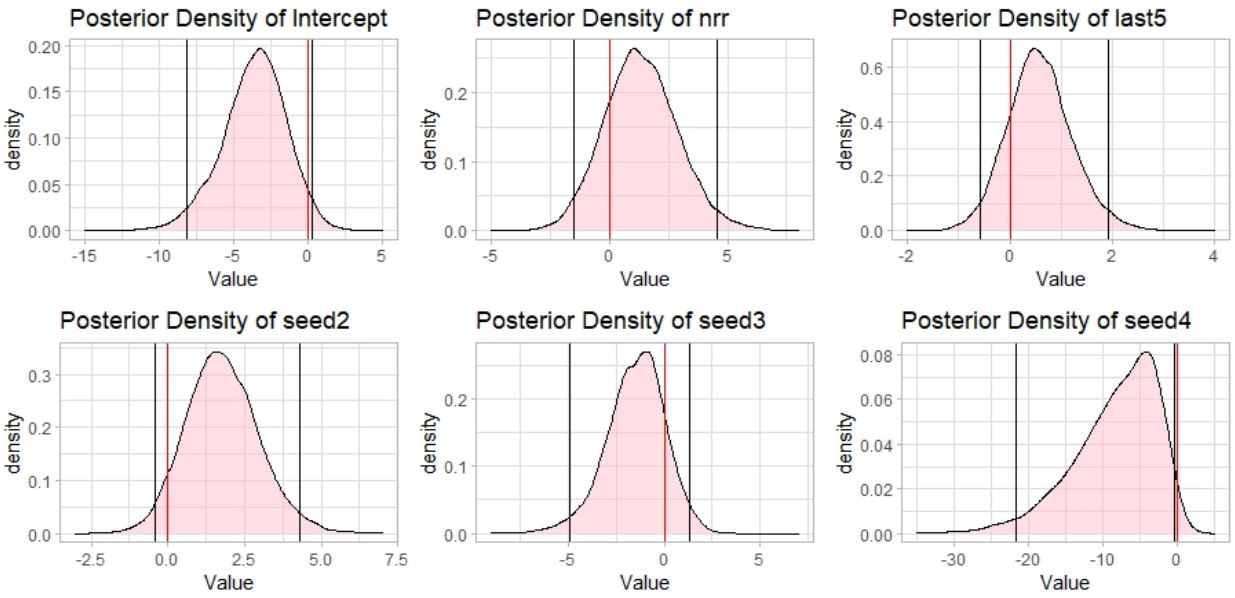

Figure 5: Posterior density plots

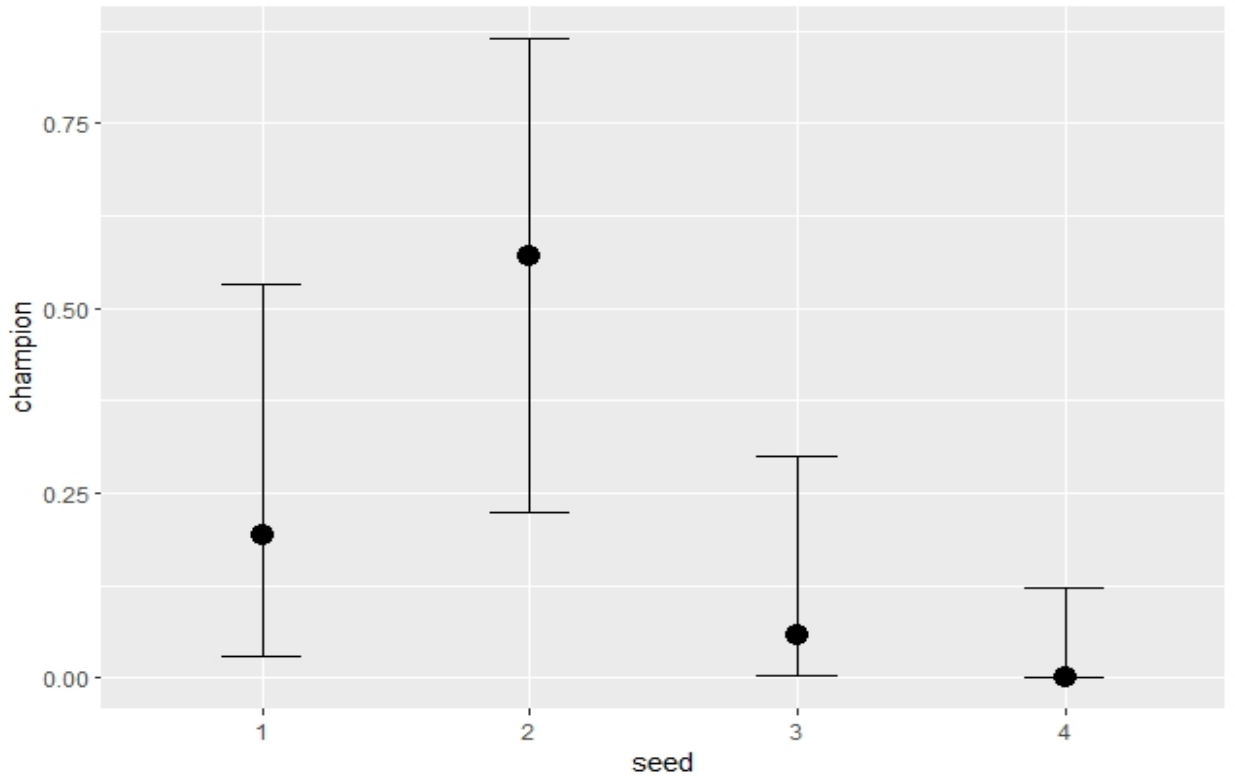

Figure 6: Conditional effect of seed variable

8 Tables

Year	result	champion	seed	nrr	last5
2017	1	1	1	0.784	3
2017	2	0	2	0.176	4
2017	3	0	4	0.641	1
2017	4	0	3	0.599	3

Table 1: League standings for year 2017

Parameter	Point estimate	Upper confidence interval
$\beta_{intercept}$	1	1.00
β_{seed2}	1	1.00
β_{seed3}	1	1.01
β_{seed4}	1	1.00
β_{nrr}	1	1.00
β_{last5}	1	1.00

Table 2: Scale reduction factor

Parameter	estimate	SE	lower CI	upper CI
$\beta_{intercept}$	-3.56	2.10	-8.19	0.37
β_{seed2}	1.76	1.20	-0.47	4.31
β_{seed3}	-1.50	1.50	-4.64	1.14
β_{seed4}	-8.23	5.51	-20.75	-0.34
β_{nrr}	1.31	1.53	-1.53	4.54
β_{last5}	0.57	0.61	-0.58	1.87

Table 3: Posterior parameter estimates (logit scale)

Parameter	2.5%	Median	97.5%
$\beta_{intercept}$	0.0003	0.0290	0.5646
β_{seed2}	0.3682	0.8507	0.9881
β_{seed3}	0.0059	0.1994	0.7974
β_{seed4}	0.0000	0.0009	0.4825
β_{nrr}	0.1708	0.7958	0.9887
β_{last5}	0.3521	0.6365	0.8581

Table 4: Posterior parameter estimates (probability scale)

Year	seed	nrr	last5	P(champ)	result
2021	1	0.481	3	0.30	3
2021	2	0.455	2	0.55	1
2021	3	-0.140	4	0.13	4
2021	4	0.587	3	0.02	2

Table 5: Prediction for year 2021