Determining Likelihood of RCB Wins in the IPL

Abdullah Farouk

ABSTRACT

We look to identify variables that explain the variation in the outcome of RCB's matches. We make use of the IPL dataset from Kaggle with data on 636 matches to do so. We employ several methods to achieve this goal. We first begin by performing a univariate analysis to identify individual predictors that we think maybe of importance. We fit a logistic regression model to our data to assess which factors best explain win probabilities. We convert our wins into counts to check if the same factors are statistically significant. Our analysis yields interesting insights about the importance of inclusion of variables like the identity of the umpire of the game.

December 14th 2017

INTRODUCTION

Cricket is one of the oldest sports ever played, with its origins dating back to the early 16th century. The game is played in three formats - Test Matches, ODIs and T20s. We focus our research on T20s, the shortest (time-wise) format of the game. The Indian Premier League (IPL) is the most-attended cricket league in the world and ranks sixth among all sports leagues. The brand value of IPL in 2017 was US$5.3 billion. As a sport, Cricket is estimated to have 1.5 billion fans worldwide. IPL betting markets are a multi-billion-dollar market. Therefore, there is a strong incentive to design models that can predict the outcomes of games and beat the odds provided by bookers. The annual cricket gambling market is thought to be worth $10 billion.

Whilst most predictive models have focused on team and opposition player statistics we do something different. We turn our attention towards variables that aren't often studied (like toss decision and game umpires). We seek to determine if variables like these play an influential role in explaining the variation in the outcome of games played by the Royal Challenger's Bangalore's (RCB) in the IPL We realize however that our data comes from an observational study and therefore our findings cannot be generalized to teams' other than RCB in the IPL.

A brief explanation of cricket

Cricket is a bat and ball game played between two teams of eleven players. The game is played on a cricket field with a rectangular 22-yard pitch in the centre. On each end is a set of three wooden stumps called wickets, with bails on top. Each phase of play is known as an innings during which one team bats while the other fields. There are two innings in T20 games. Each team swaps roles at the end of an innings. The winning team is usually the one with the most number of runs scored, unless in the event of a draw. In this case both teams have the same run totals.

At the start of the match, the two team captains flip a coin to decide who will bat or field first. The game begins when a member of the fielding team delivers (bowls) the first ball of the first over. An over is a set of six deliveries by the same bowler. One way a batsman loses his wicket batsman (gets "OUT") is when the ball hits the wickets directly and dislodges the bails. The main objective of the batsman is to protect his wicket and score as many runs as possible, generally speaking. The bowler's objective is to prevent the scoring of runs and dismiss the batsman. Adjudication is performed two on field umpires.

3.1 References

1)" Predicting the Match Outcome in One Day International Cricket Matches, while the Game is in Progress" (https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3861745/). In this paper, the author uses multiple regression analysis to predict team totals and determine the outcome of ODI's

2) "Fitting of Logistic Regression Model for Prediction of Likelihood of India Winning or Losing in Cricket Match" (http://www.icmis.net/icmis15/icmis15cd/pdf/S5044-final.pdf) The author uses a logistic regression model to determine the likelihood of India beating Australia in an ODI

3.3 Data Description

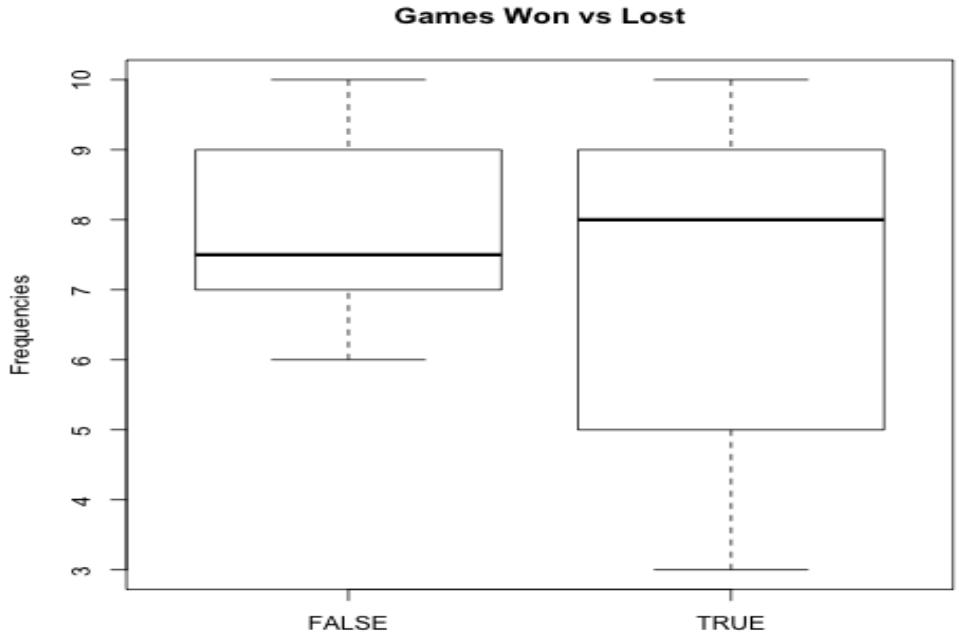
The dataset titled "Indian Premier League" on Kaggle's website was used for the purposes of this report. It contains seventeen variables and information on 636 games, ten teams and ten seasons. The data is longitudinal in nature, obtained by forming multiple observations on each team over time.

The table below provides information on all the variables used in this report

Variable Name	Description
Id	Integer; Unique identifier of game.
Season	10 level factor; Year game was played in
City	31 level factor; Name of Indian city game took place in
Date	450 level factor; Specific date of each game (in yyyy/mm/dd format)
Team1	14 level factor; Name of first team
Team2	14 level factor; Name of second team
Toss_winner	14 level factor; Name of team that won the toss
Toss_decison	2 level factor; Decision to either bat or field
Result	3 level factor; Game tied, winning team scored more runs or no result
dl_applied	Integer; method to determine target score for team batting second to achieve in the event of game interruption by weather conditions
Winner	15 level factor; Name of winning team in game between team 1 and 2
Win_by_runs	86 level factor; Number of runs opposing team won by
Win_by_wickets	11 level factor; Number of wickets left
Player_of_match	202 level factor; Name of player with best in game performance
Venue	45 level factor; Name of stadium
Umpire1	45 level factor; Name of first umpire
Umpire2	46 level factor; Name of second umpire
Toss_outcome	2 level factor; RCB chose to either bat or field when they won the toss
Opponent_faced	14 level factor; indicating name of team RCB played against
Win_streak	2 level factor; 1 if RCB won the game; 0 otherwise
win_count_1	Number of games RCB won each season of the IPL

EXPLORATORY DATA ANALYSES

We provide some preliminary findings in this section. To begin, we note that 152 of the 636 games in our dataset (24%) were played by RCB. Focusing our attention on them from now on, we notice interesting patterns in our data. They won 48% (73 of 152 games played) and lost 52% of them. A preliminary analysis of their wins over time (denoted by the boxplot corresponding to true) show some interesting findings. They have both won and lost a maximum of 10 games in a given season. However, their wins seem to be more volatile. They seem to have a median loss of about seven games a season. This is lower than their median number of wins. Of the 73 games they won, they won 31 of them with no wickets in hand, i.e. all batsmen were "out".



In the diagram above on the x axis False represents games lost whilst True represents games won.

We now analyze certain variables of interest. Through this we hope to form an understanding of the factors that influence RCB's chances of winning. We look at three primary variables. They are:

- Toss Decision (to bat or field upon winning the toss)
- City game is played in (Home ground advantage)
- Opposing Team

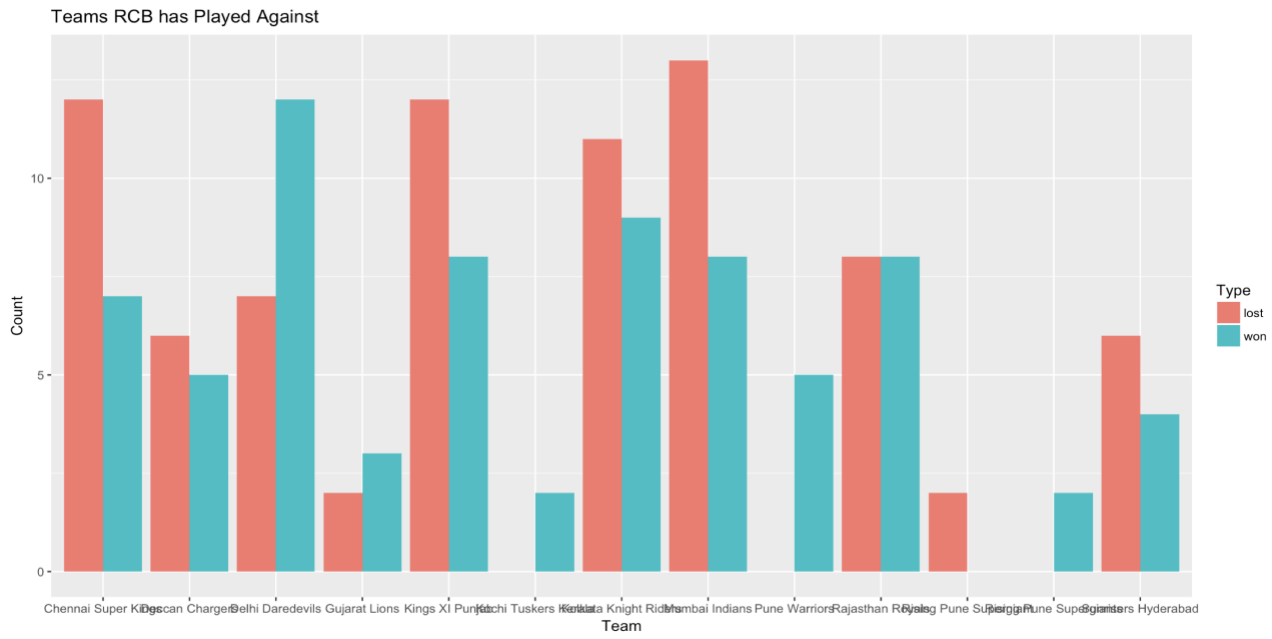
Toss Decision: -

They won the toss in 70 of the 152 games they played. They won 52% of the games they chose to field in and 45% of the games they chose to bat in. The data suggests a preference for fielding first, with them having chosen to do so in 50 of the 70 games they won the toss in.
We ask ourselves if fielding first leads to a higher win rate. To check this, we employ a chi square test of differences. We are interested in determining if the difference (between choosing to bat or field) has a significant effect on the outcome of the game (win vs loose) The value of our chi squared statistic is 0.28 with 1 degree of freedom with a p value of o.59. Thus, we find insufficient evidence to support the claim that the difference is significant.

City game is played in:

It is a popular belief among sports fans that playing on home grounds should result in higher wins, given the immense support by the crowd. To test this hypothesis, I look at RCB's wins, across the different cities they have played in. We notice that they played the most number of games (62 out of 152) in Bangalore. Contrary to popular belief, we find that RCB only won 48% of the 62 games they played at home. This seems to indicate that playing at home is more likely to have an adverse effect on RCB winning. To test if the city a game is played in leads to a significant difference between winning and losing, we employ a chi squared test once again. It has a value of 20.42, 24 degrees of freedom and a p value of 0.67. Hence, we find insufficient evidence to reject the null at any reasonable significance level.

Opponent faced:
Here we asses if RCB's odds of winning are significantly better against specific teams than others. The graph shows us some interesting trends. We notice that it suffered its largest cumulative loss to the Mumbai Indians (MI) whilst garnering its largest cumulative win against the Delhi Daredevils. Does this violate the notion that RCB has an equal chance of winning or losing to either one of the teams? Under the null, we expect a non-significant difference between the number of wins and losses against a specific team. To test this, we compute a chi squared statistic. It has a value of 16.302 and a p value of 0.178. Thus, we find insufficient evidence to support the claim that the difference is significant at a reasonable significance level.



Our initial analysis revealed some interesting variables whose effects we would like to asses. They include RCB's choice to field when winning the toss, playing a game on home grounds and the effect of playing a game against MI. Though our initial tests do not provide any statistical evidence to indicate their significance, we study them due to the many times RCB has chosen to field, played in their home grounds and lost to MI despite playing against them the most.

Confirmatory Analysis

In this section, we attempt to identify statistically significant variables that affect the outcome of a game RCB is playing in. In addition, we hope to use the models we fit to verify our findings from our univariate analysis described in the previous section. Our main objective is to identify variables that can accurately capture RCB's chances of winning. We do so by:

1) Fitting a Logistic Regression Model:

We convert the winner variable in our dataset into a binary variable called win_streak. It takes a value of 1 to denote a win by RBC and 0 to indicate a loss by them.

2) Fitting a Poisson Regression Model:

The response variable win_count_1 represents the number of wins in a season. The table below provides an overview of the number of games RCB won and lost during a particular season.

Season	Lost	Won
2008	10	4
2009	7	9
2010	8	8
2011	6	10
2012	7	8
2013	7	9
2014	9	5
2015	8	8
2016	7	9
2017	10	3

Data Preparation

Before we proceed on to our analysis, we mention the following manipulations we carried out on the variables in the dataset. We did so for the following reasons. Some variables such as team1 and team2, for instance, are highly correlated. This leads to convergence problems and quasi perfect linear separation when trying to fit a quasi-binomial model to the dataset.

The next big issue we tackle is the large number of sparse categories within most of the categorical variables. A good example would be the variables win_by_wickets and city. Both of these variables have about 10 levels but only about 3 levels contain most of the observations. The remaining contain a tiny proportion of the observations. They represent rarities. To deal with them, we lump the infrequent factors and focus our analysis on concentrated levels.

In addition to the above we define new variables to capture the effects of interest in our model. The first is a factor of names of the opposing team. The next one is called toss outcome. It is the decision by RCB to bat or field in the event of winning the toss. We find it meaningless to include toss_winner and toss_decison in their given form as variables in the model. This is because we are trying to assess whether RCB's decision to bat or field results in a higher chance of winning. When RCB loses a toss they cannot choose to bat or field. We circumvent this by creating an interaction term to take the place of these two variables in the dataset.

We then proceed to check for correlations amongst variables. Our results reveal that city and venue, date and season and win by run and win by wickets are highly correlated. It also suggests that date and season are correlated (significantly) with two other variables. Thus, we eliminate date, season, venue and win by runs from the dataset to avoid convergence issues. Given the large number of explanatory variables we have, the glm algorithm still fails to converge. To overcome this, we employ two strategies of dealing with our categorical variables.

1) Converting them into integers: -
Under this strategy we replace the data for a particular observation with the integer representing its level. This alleviates the problem of convergence, though it gives rise to a new problem. How do we interpret this variable? It is also not clear if the number accurately reflects the difference in levels (for instance, do we really expect the distance between winning by 0 wickets and 3 wickets to be the same as that between 6 and 9 wickets).

2)Converting them into dummy variables:
To overcome the issues mentioned above, we convert them into dummy variables. Not only is their interpretation relatively simple, it captures the effect of each level of each of our explanatory variables. This makes it easier for us to understand and interpret significant effects.

We carry out our analysis using both methods. They identify the same effects as significant. This is good as it helps verify our findings. However, we choose to only report our findings from our use of dummy variables as they are easier to interpret.

1) Fitting a Quasi Binomial regression model:

Since our response variable is now binary we can analyze it using a binomial error distribution model. Before we fit a logistic regression model, we check for dispersion in the data. In the presence of dispersion, we are likely to observe inflated standard errors and the binomial model is more likely to identify effects as significant. To avoid misidentification of significant effects, we fit a quasibinomial model. The dispersion parameter has a value of 1.128309. This indicates over dispersion within the data. Thus, we decide to proceed with the use of a quasibinomial model.

We first regress all the variables in the dummy variables dataset (rcb_dummy) created against win_streak to determine which ones are statistically significant in explaining the observed variation. To do this, we use a chi square test statistic and sequentially drop variables to identify important variables. It identifies player_of_matchJH Kallis as significant at the 10% level and player_of_matchOther as significant at the 0.1% level.

To test the hypothesis that playing a game in Bangalore, choosing to field when winning the toss and playing against the Mumbai Indians has no effect on the log odds of RCB winning, we use a likelihood test. The test compares the model with just player of the match as an explanatory variable, with the one described above. Since we are conducting a multiple hypothesis test, we use an F statistic. The value of the F statistic is 0.8257 and it has a p value of 0.4817. Hence, we find insufficient evidence to reject the null at any reasonable level of significance.

Our results are consistent with our findings from the analysis conducted preciously. None the less, we keep the variables of interest to us, whose effects are not statistically significant, and add player of the match other to it to form our final model for modelling the odds of winning.

Our results suggest that the player of the match is an important factor in determining the odds of RCB winning the game. We don't quite understand why might this be the case? To do so we first identify the effect it captures.

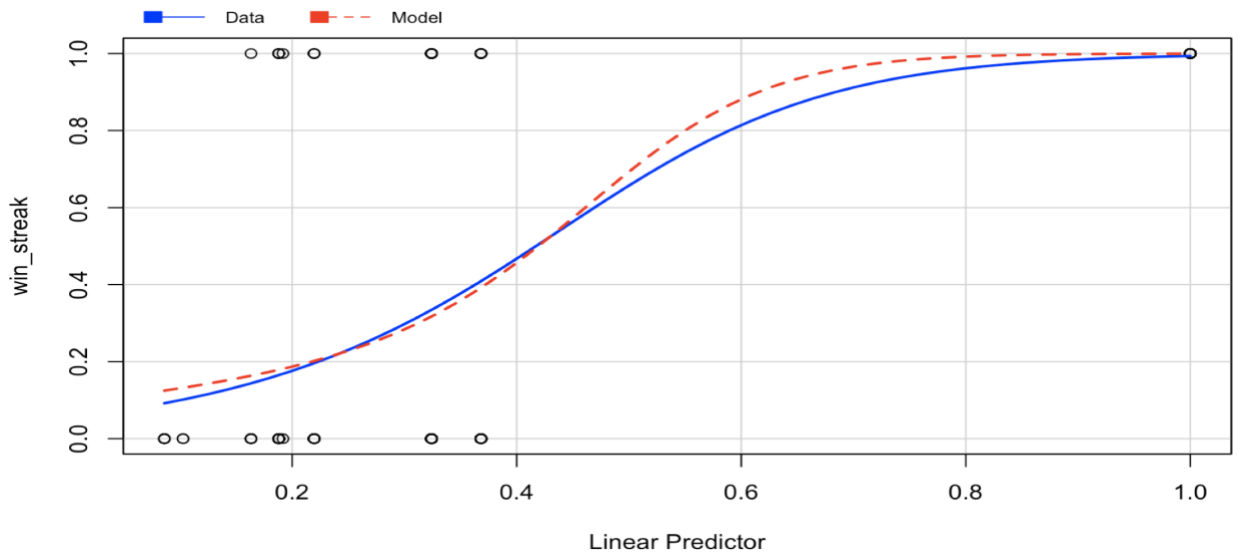
The player of the match is ***usually*** awarded to a member with the most outstanding performance or impact in the game. Players who tend to win it the most are those who have been performing well consistently throughout a season. Hence if players who are members of the opposing team win, it is a good indication that they and subsequently their team are likely to have been performing well throughout a season.

The variable player of match other though statistically significant, is not very meaningful. This is because we previously lumped all sparse levels of it together. Therefore, to understand it better we examine the original levels of the variable player of match. We observe that 63 players were named player of match when RCB lost a game. Of them, 89% (56 out of 63) were members of another team. We also note that RCB has lost 52% (79 out of 152) of its games in the IPL. We believe that this historic loss record combined with the fact that 89% of non RCB team members won the award leads our model to identify this variable as having an adverse effect on the odds of RCB winning. This is because a large number of wins by a specific player of the opposing team is a good signal of that particular team's good form and more of a challenge to defeat.

We also note that Kallis used to be a part of RCB, during which time he won the award 5 times. He then switched to the Kolkata Knight Riders, against whom RCB lost to 55% of the time (11 out of 20 games) with whom he won it once again. Hence, we find evidence that supports the key role his performance played in RCB winning games. We believe this is why his switch negatively impacts their chances of winning. We find it remarkable that this effect is captured by our fitted glm model! None the less we choose to leave this effect out of our model as he has now retired from playing the sport.

In order to be confident in the findings from our model, we carry out various model diagnostics tests. We examine plots of residuals for each of our explanatory variables. They help us identify if there are any nonlinear relationships we might have to account for. We do not see any need for one. Furthermore, all our explanatory variables are dummy variables and none of them are continuous. Hence, we are confident in the form in which our explanatory variables have been incorporated into our model.

The next step we undertake is to check for influential points in our model. If removal of an observation causes substantial change in the estimated coefficients, we consider it influential. Cook's distance is a summary measure of influence. A large value of Cook's distance indicates an influential observation. Observations with large studentized residuals are likely to be outliers whilst those with a large value of cook's distance are most likely to be influential. We find observation 91 and 128 to be influential (large cooks distance and studentized residuals).
Thus, we refit our model without these observations. A plot of our fitted model is shown below.



The figure above helps visualize how well our model fits the data. It plots predicted values of RCB winning a game for a linear combination of the explanatory variables in our final model. It displays the odds of winning on the vertical axis and a linear combination of the explanatory variables on the horizontal axis. It contains a scatter plot of the two variables, a smooth fit function for the variables in the plot ("Data"), and a function that displays the predicted values as a function of the horizontal axis variable ("Model"). The similarity between the two graphs indicates that our model fits the data well.

Updated Coefficients for Quasi Binomial Model				
	Estimate	SE	Estimate 2	SE 2
(Intercept)	3.372	0.827	20.046	1328.151
cityBangalore	-0.47	0.479	-0.728	0.419
tossfield	0.312	0.471	0.195	0.399
`opponent_facedMumbaiIndians`	-0.673	0.726	-0.899	0.675
player_of_matchOther	-4.246	0.807	-20.781	1328.151

We notice that almost all of our coefficients change. They have all increased in magnitude but their signs remain the same, retaining the interpretation of their effects on the odds of winning as mentioned above.

Our model tells us that RCB's odds of winning a game, provided we hold all other explanatory variables in the model constant:
1) Decreases by 0.48 exp(-0.728) if the game is played in Bangalore
2) Increase by 1.22 exp(0.195) if they choose to field when they win the toss
3) Decreases by 0.41 exp(-0.899) if their opponents are the Mumbai indians
4) Decreases by 0.009 exp(-20.781) if player of the match is other (represents players who were awarded this less frequently and are now all lumped together)

Our dichotomization of the winner column in our dataset, provided us with interesting insights into variables that affect the odds of RCB winning a game. To verify our findings, we ask ourselves if the same variables will be significant if we change our response variable?

1) Fitting a Quasi Poisson Regression Model:

Since our response variable is now a "count" of wins, we feel it is justifiable to try and fit a Poisson error distribution model. One of the assumptions of a poisson model is that the conditional variance equals the conditional mean. To check if this assumption is violated we fit a quasi-poisson model. The value of its dispersion parameter is 0.626. This indicates our data is under dispersed and suggests that a fit based on the poisson model would be inappropriate. Hence to proceed, we use a quasi-poisson model.

We first regress all the variables in the dummy variables dataset created against win_count. To do this, we use a chi square test statistic and sequentially drop variables to identify ones that are statistically significant. Whilst the initial regression on all variables finds cityBangalore significant at the 5% level, our likelihood test does not. It identifies `umpire2SJA Taufel`, `umpire2S Ravi` and player_of_matchOther as significant at the 5% level and `umpire2RJ Tucker` and `umpire2RB Tiffin` significant at the 6% and 9% level respectively.

We use a likelihood test to assess the effects of the variables of interest to us on the expected log count of wins of RCB. Since we are carrying out a multiple hypothesis test, we use an F statistic. The test compares the model with the variables we mentioned it found significant, with a model that includes them and our variables of interest. The value of the F statistic is 0.4773 and it has a p value of 0.6986.

We find insufficient evidence to reject the null, that playing a game in Bangalore, choosing to field when winning the toss and playing against the Mumbai Indians has no effect on the expected log count of wins, at any reasonable level of significance. Though we don't find any statistical evidence, we combine our variables of interest with those found to be significant to formulate a final model to explain variation in win rates across the different seasons.

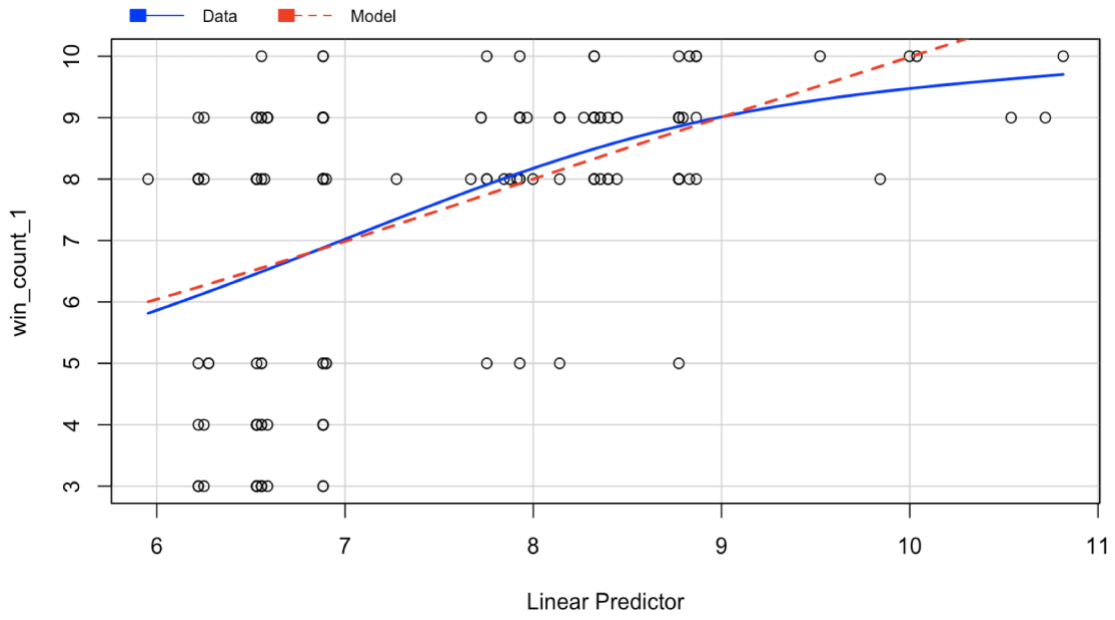
In order to verify the validity of our findings we carry out some model diagnostics. We first look for influential observations. We find that observations 4, 48 and 145 are influential (large cooks distance and studentized residuals). We remove those to asses if there are significant changes in our coefficients. The only difference we notice is that our coefficients increase, in an absolute sense, in magnitude.

Updated Coefficients for Quasi Poisson Model				
	Estimate 1	SE 1	Estimate 2	SE 2
(Intercept)	2.14811	0.05075	2.17172	0.04992
`opponent_facedMumbai Indians`	-0.00794	0.06861	-0.04383	0.06822
tossfield	-0.03915	0.04966	-0.04868	0.04814
cityBangalore	-0.04169	0.04814	-0.05262	0.04718
`umpire2RB Tiffin`	0.13938	0.0935	0.18332	0.09289
`umpire2RJ Tucker`	0.06103	0.10411	0.05507	0.10057
`umpire2SJA Taufel`	0.25357	0.08728	0.25308	0.08436
`umpire2S Ravi`	0.17369	0.08558	0.16766	0.08265
player_of_matchOther	-0.22962	0.04945	-0.24261	0.04833

Our model provides us with interesting findings. It tells us that RCB's expected win count should, provided we hold all other explanatory variables in the model constant:

1) Decrease by 0.95 exp(-0.05262) if the game is played in Bangalore
2) Decrease by 0.95 exp(-0.04868) if they choose to field when they win the toss
3) Decrease by 0.95 exp(-0.04383) if their opponents are the Mumbai Indians
4) Decrease by 0.78 exp(-0.24261) if player of the match is other (represents players who were awarded this less frequently and are now all lumped together)
5) Increase by 1.28 exp(0.25308) if Taufel is umpiring the game
6) Increase by 1.18 exp(0.16766) if Ravi is umpiring the game
7) Increase by 1.17 exp(0.15998) if Tiffin is umpiring the game
8) Increase by 1.06 exp(0.05507) if Tucker is umpiring the game

The plot below displays the count of wins in a given season on the vertical axis and a linear combination of the explanatory variables on a horizontal axis. The explanation of the different lines remain the same as defined in the plot under the quasi binomial fit. Thus, given how similar the models are we feel that our specification fits the data reasonably well.



Modeling wins as counts per season reveals an interesting significant variable. It says that the identity of umpire2 of a game is statistically significant. A cricket umpire is a person who has the authority to make decisions regarding the game while it is being played. A game generally has two (on field) umpires who rotate positions and roles throughout the game. They perform the same functions, more or less, during the game. Their decisions can have a significant impact on the outcome of a game. Hence to understand the effect of umpires in the context of RCB's win count better, we look at win-loss records under the different umpires.

We notice that, with the exception of Taufel, RCB has won as many games, if not more, than they have lost under each umpire. Thus, it seems reasonable that the identity of the second umpire has a positive effect on the expected log count of wins.

An interesting observation is that RCB’s choice to field upon winning the toss increases the odds of winning a game. The same decision though leads to a decrease in the expected count of wins, holding all other explanatory variables constant. We believe this is a reflection of the fact that they have only won 52% of the games they chose to field first in, despite having chosen to do so in 71% (50 out of 70) of the games they won the toss in. This suggests that fielding first improves the odds of winning a game but is likely to result in fewer cumulative wins in a given season.

Conclusions and Limitations

Our findings from fitting glms are consistent with our univariate analysis. In addition, we find that the identity of the second umpire is significant when modelling the number of games RCB wins in a season. The variable player of the match other (a measure of outstanding performance by an individual player) is the only variable that is statistically significant under both models. None the less, we choose to include the following variables in our final model as we believe they are most useful to include in any model that attempts to predict the chances of RCB winning a game. They are:

1) If the game is played in Bangalore
2) If they won the toss and choosing to field
3) If their opponents are the Mumbai Indians
4) If the second umpire is either Ravi, Taufel, Tiffin or Tucker
5) Who has been winning the player of the match award the most over a given period of time

One of the biggest advantages of the models we present in this report is the treatment of our explanatory variables as dummy variables. This allows us to identify the specific impact of a particular level of our variable of interest. However, because we only have 152 observations on RCB, it is difficult to represent all the levels of each variable in our model without running into issues of quasi linear separation or lack of convergence. Thus, we essentially miss out on capturing the effects of sparse categories whose impact maybe me useful to understand for predictive purposes (like in determining the outcomes of a final that RCB is a part of).

The other limitation that must be mentioned is our construction of our count response variable. Since we have longitudinal data, an alternative way to compute the number of wins in a given season would be to use a rolling window to compute the cumulative sum of wins over a given season. However, we were unable to build a working code to implement this idea.

One final point that must be mentioned is with regards to the longitudinal nature of our data. We have both variation across time and across teams (observational units) in our data. We feel that a better way of accounting for this variation would be by fitting a mixed effects model or a general estimating equations model to the data. Our knowledge of such models is very limited at this point in time and therefore we have refrained from using these tools in our analysis.

On a final note, we reiterate that our data comes from an observational study and therefore our findings cannot be generalized to teams' other than RCB in the IPL.