Sabermetrics: The Effect of Pitching and Runs on the Probability of Winning

Benjamin Ryu[*]

May 19, 2024

Abstract

This article attempts to estimate the effect of the ERA on the probability of winning. We ran Logistic and Linear Probability models to infer our chance of winning the division. We test for variable inflation factors to ensure we are able to make a causal claim to our result. Overall we Believe the ERA out of all the indicators we have tested for has the most significant effect on the chance a team wins the division. This supports previous intuition about baseball. The value of a pitcher has on the team is significant when it comes to winning.

Keywords: Baseball, Sabermetrics

[*] Department of Economics, San José State University

1 Introduction

Baseball is at the forefront of metrics in sports because of the availability of data for empirical analysis, which is unlike any other sport. With the rise of metrics and sports forecasting models, we can predict outcomes of games, player salary, and other sports statistics to determine how a team may spend their money during free agency.

Our primary interest is if we can build a model that can predict a team's chances of winning the division and by extension being able to play the post-season based on team statistics available from Major League Baseball(MLB). Currently one of the most popular ways to measure a field player's contribution is Win Above Replacement(WAR) calculated by taking a player's runs above average a player is worth in his batting, base running, and fielding + adjustment for position + adjustment for league + the number of runs provided by a replacement-level player divided by runs per win. For pitchers, WAR is calculated differently and is based on the inning pitched total. This is measurement is only attributed to the individuals and not a team statistic. Although WAR is a great statistic, unfortunately in the scope of this paper we are not interested in it. We are more interested in team statistics, as baseball is a team sport.

For our article, we are going to look at the Earned Run Average(ERA) as our primary statistic, we will also be testing other statistics. We aim to find the degree to which ERA contributes to a team win and by doing so we can better understand what teams should value in re-building periods to have a successful season. The probability of winning beyond just the division such as the World Series is more determined by the series of 7 games and will not be as affected by the season-long statistics in the same way it does for division wins. This is because a team may perform poorly during the season but suddenly improve during the post-season 7-game series, therefore any statistic they built will not reflect their probability of winning that series.

Why are we primarily interested in ERA? ERA is typically accepted as a good way to measure a pitcher's ability, and in baseball pitching is one of the most important roles.

So how does baseball work? For people to understand our study we don't need a deep explanation of how the game of baseball works. The most important information that is needed is that there are batters and pitchers in a baseball team, these two roles are the most important when it comes to directly affecting how points are scored. A team that outscores their opponent in a game is considered a winner and to be a division winner they have to win more games than their opponent over a season, in which 162 games are played. So when a team's pitching ERA is low, this indicates that the team does not allow a lot of runs or points scored on them. Another statistic that's directly involved is runs scored, so what are runs scored? runs scored put simply is an indicator of the total run or point scored in the season. Therefore we are most interested in ERA and runs scored compared to all other statistics, which are secondary as it does not directly impact points except for home runs.

Our main method will be using logistic regression to find the probability of going to the off-season with a two-way fixed effect. This will allow teams to determine better what statistics to look for when scouting for new players as each player's statistics are available to team scouts. And hopefully, contribute to a more even playing field for smaller market teams.

2 Literature Review

Albert (2010) is an overview of the sabermetrics throughout the years and its use in baseball. Albert (2010) goes into how these statistics are calculated breaking down some of our core statistics such as On Base Percentage(OBP) and Earned Run Average(ERA). Albert (2010) continues to explain that ERA might not be the greatest measurement of a pitcher's ability and only 9 percent variability can be explained by ERA. A lot of factors have been in consideration on the result of the player ERA but we are not particularly interested as we look at team ERA rather than individual ERA and through the law of large numbers we are hoping to be close to the true mean.

Lindsey (1959) looks into the use of statistics for the operation of a baseball team. Lindsey (1959) looks into a player's individual record to make on-field decisions. example were given such as when the starting pitcher rotation and relief pitchers. Lindsey (1959) continues to state that batter average sampling error is too great to make a useful analysis. Also, cite that the statistics are important in making pitcher rotations and relief pitching decisions. Although not direly not of use, it does provide insight into how statistics are used in baseball.

Cover and Keilers (1977) paper might not be in direct relation to our subject but it offers an interesting perspective on baseball statistics. Cover and Keilers (1977) creates a new statistic called offensive earned-run-average which can be thought of as an inverse of earned-run-average where instead of the pitcher earned runs the batter earned runs. For the batter the higher the OERA is the better while for pitchers they aim for a lower number of earned runs averaged. This does show us that we can use observable data to better determine a player's input by creating new statistics. An example of modern statistics used from mathematical calculation of observable data is On Base Percentage plus Slugging.

Houser (2005) paper is the most recent and relevant paper out of our literature review. Houser (2005) found that the on-base percentage has been the most important statistic, which can be seen in Billy Beane's team during his time at A's. Houser (2005) continues to reveal that Walks and Hits per inning and on-base percentage were the most important statics in making personal decisions. You will see we came to a similar conclusion on how the team should be built.

3 Data

Our primary data source is derived from Lahman's Baseball Database which derives baseball statistics from the Major League Baseball(MLB). The data is collected from 1998-2019 season giving us a total of 22 seasons and 30 teams, which comes to a total of 660 observations. We have omitted the 2020 season as it was a shorter season due to the effect of the COVID-19 pandemic. Although there are mountains of baseball statistics that are gathered throughout the season the most important statistics in determining wins are the runs scored and Earned Run Average(ERA). The objective of each team is to score more runs than they allowed. All these statistics are either directly or indirectly related to runs or points scored. Therefore, one of the statistics often looked at is a pitcher's ERA. The ERA is the amount runs a pitcher lets the opponent team make per 9 innings, additionally, an average game runs for 9 innings. Next would be the run scored by each team, theoretically the higher this number is the more they have scored and the more likely they outscore the opposition but this statistic only matters if you have a low ERA as it is possible to score a high number of runs and also allow a high number of runs, this will essentially balance out and decrease their chance of winning the division compared to if they had a lower ERA and a higher run scored. These two statistics would in theory work in tandem in building a winning team. All other statistics in baseball are typically more important on an individual level and on-field decisions.

Table 1: Summary Statistic

Statistic	N	Mean	St. Dev.	Min	Max
Run Scored	660	746.044	85.227	513	1,009
On Base Percentage	660	0.328	0.015	0.292	0.373
Homeruns	660	173.744	36.756	91	307
Earned Run Average	660	4.277	0.543	2.940	6.010
Walks And Hits Per Inning Pitched	660	1.363	0.097	1.099	1.705
Teams	660	-	-	0	30
Years	660	-	-	1998	2019

Next, we look at Walks and Hits per Inning(WHIP) and On Base Percentage(OBP). These statistics are not observed statistics but rather a calculation of other statistics. For WHIP we take the sum of walks allowed and hits allowed divided by inning pitched. While for OBP we take the sum of hits, hit-by-pitchs, and walks divided by the sum of at-bats, hit-by-pitch, sacrifice files, and walks. Therefore we can eliminate any statistic that adds up to WHIP and OBP in our models to avoid multicollineary.

4 Empirical Methods

We have 3 equations for the article, Equation(1) is our logistic model that takes both the offensive baseball statistic $Batter_{it}$ and defensive statistics $Pitcher_{it}$ and fixes for variance in year b_t and team a_i.

$$Pr(Y = 1|Batter_{it}, Pitcher_{it}, a_i, b_t) = F(\alpha + \beta_1 Batter_{it} + \beta_2 Pitcher_{it} + a_i + b_t) \tag{1}$$

$$Logistic = \frac{1}{1 + e^{-(\alpha+\beta_1 Batter_{it}+\beta_2 Pitcher_{it}+a_i+b_t)}} \tag{2}$$

$$LinearProbabilityModel = \alpha + \beta_1 Batter_{it} + \beta_2 Pitcher_{it} + a_i + b_t \tag{3}$$

Equation (2) gives us the logistic model's probability density function. Equation (3) is our Linear Probability Model in which we can infer the magnitude of our variable effect on winning the division, which can't be measured in the logistic model without further calculation. We are expecting that Equation (3) is not going to be as robust as our logistic model. We are also going to be running this model with individual statistics. We also going to run a Variable Inflation Factor to help us to a more robust result.

5 Results

Our results are varied depending on the model and regression used. Table 2 is our main regression table that runs a logistic regression on the division win by testing multiple different covariates. We also ran each covariate on its own in case of multicollinearity.

Table 2 has 8 models in total, the first model runs all our variables of interest on division win and we see that runs scored, OBP, ERA, and constant are significant while all other variables are not. We also see that ERA and the constant are the only negative variables, which makes theoretical sense as a lower ERA means a higher probability. The magnitude of the model can not be spoken of as of now until we run a linear probability model. When

Table 2: Logistic Regression: Division Win

	Dependent variable:							
	Division Win							
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Run Scored	0.016**	0.015***				0.022***		0.027***
	(0.006)	(0.005)				(0.003)		(0.003)
On Base Percentage	68.518**	41.752					115.096***	
	(31.398)	(25.399)					(14.306)	
Homeruns	0.009	0.004						
	(0.009)	(0.007)						
Earned Run Average	−5.030***		−2.790***		−3.709***			−4.612***
	(1.028)		(0.775)		(0.429)			(0.578)
Walks And Hits Per Inning Pitched	1.855		−5.850	−19.280***				
	(5.202)		(4.204)	(2.273)				
Constant	−20.593**	−29.039***	18.405***	25.060***	14.202***	−19.825***	−40.670***	−4.461
	(8.871)	(6.231)	(3.820)	(3.417)	(2.256)	(2.533)	(5.021)	(3.409)
Observations	660	660	660	660	660	660	660	660
Log Likelihood	−163.627	−219.835	−216.355	−223.295	−217.331	−221.205	−230.069	−166.099
Akaike Inf. Crit.	447.253	555.670	546.709	558.591	546.662	554.411	572.138	446.197
Note: Fixed for Years and Team								*p<0.1; **p<0.05; ***p<0.01

we run only the batter statistics model (2) we see that run scored and constant remain significant with no sign change, while OBP becomes insignificant. In our pitcher statistic model (3) we see that the ERA and Constant remain significant with no changes in sign but the WHIP does change signs to our expectation. when variables are run alone on division win we see that the statistic changes into higher statistical significance. Model (8) is our most important model as it takes one offensive statistic and a defensive statistic to determine the probability. Later in Section 5.1, we will understand why although all other models show variables that are significant and even control for another variable, model 8 holds out the most relevance.

Our Table 3 compares the logistic model with the Linear Probability model(LPM), we are able to read the magnitude of the LPM, unlike the logistics model. Although the results on LPM are interpretable the robustness of the results is not comparable to the Logistic model and should not be taken as a casual result. We do see changes in significance from our logistics to our LPM; OBP and the Constant become statically insignificant while run scored and ERA remain statistically significant as in model (1) to model (5). The sign

Table 3: Division Win: Logistic versus Linear Probability Model

	Dependent variable:							
	Division Win							
	logistic				*normal*			
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Run Scored	0.016**	0.015***		0.027***	0.002***	0.002***		0.002***
	(0.006)	(0.005)		(0.003)	(0.001)	(0.001)		(0.0002)
On Base Percentage	68.518**	41.752			2.030	3.330		
	(31.398)	(25.399)			(2.669)	(2.838)		
Homeruns	0.009	0.004			0.0001	0.0001		
	(0.009)	(0.007)			(0.001)	(0.001)		
Earned Run Average	−5.030***		−2.790***	−4.612***	−0.294***		−0.273***	−0.313***
	(1.028)		(0.775)	(0.578)	(0.078)		(0.082)	(0.034)
Walks And Hits Per Inning Pitched	1.855		−5.850		−0.114		−0.617	
	(5.202)		(4.204)		(0.446)		(0.471)	
Constant	−20.593**	−29.039***	18.405***	−4.461	−0.455	−2.465***	2.183***	−0.096
	(8.871)	(6.231)	(3.820)	(3.409)	(0.740)	(0.650)	(0.397)	(0.282)
Observations	660	660	660	660	660	660	660	660
Log Likelihood	−163.627	−219.835	−216.355	−166.099	−180.353	−223.347	−223.301	−180.724
Akaike Inf. Crit.	447.253	555.670	546.709	446.197	480.706	562.695	560.601	475.448

Note: Fixed for Years and Team *p<0.1; **p<0.05; ***p<0.01

remains consistent through both our models except WHIP becomes negative in the LPM as seen in model (1) and model (5). We see that a 1 unit increase in ERA will result in a decrease in the probability of winning the division by 0.294 percentage points as shown in model (5) but in model (7) where it only pitches, we see that it decreases in magnitude to a 0.273 percentage point. In our model (8) we see a similar magnitude of change, where an increase in ERA causes a 0.313 percentage point decrease in winning the division.

5.1 Results: Variable Inflation testing for Multicollinearity

We test for Variable Inflation Factor or VIF as many of our variables are highly correlated from one another and we wanted to avoid multicollinearity. For example, a low ERA and low WHIP measure the ability of pitcher skills and will have a high VIF as seen in Table 7 with both VIf being over 4. This is because both of these statistics are related to one another, a low ERA would result in a low WHIP in most cases. We found that our VIF is only lower than 4(as seen in Table 7-11) when we run the variables or when the statistic is one offensive(batter) and one defensive(pitcher).

So why do we think there is a high VIF for most of our variables except for Table 11(Table 2 model (8))? A higher OBP will result in higher runs scored, every homerun is counted as runs scored. This is why when we put batter statistics together they will cause a higher VIF and multicollinearity. That works the same way for pitcher statistics, ergo when we only have singular statistics from two related variables such as one pitcher and one batter statistic we will undeniably see a lower VIF as evident in Table 11.

6 Discussions

ERA has the greatest impact on our models even when looking at Figure 1 where the models are fitted is presented. Only ERA and runs scored are statically significant in winning the division as seen in Table 2 models (1), (2), and (3). However, these results fail to score under 4 in VIF and are seen as noncausal due to the multicollinearity of covariates. But theoretically, this makes sense that ERA and runs scored are significant as these covariates are the most direct determinate of scores ran or prevented. Because of our worries regarding multicollinearity, we modeled Table 2 model (8) which has a lower VIF score with slight worry about multicollinearity. The reason we chose these statistics is that when we controlled for covariates they were the only ones that remained statically significant and from "ball knowledge" or general game knowledge we undertake how these stats impact results. As mentioned before they are directly responsible for the game's results.

It should be also noted that the marginal return in one unit increase in any of these variables is not directly comparable as the run scored is not accounted for by game or inning while ERA is. The average run scored in our data is reflected as 746 while the average ERA is 4.27. This also shows why even though the run score magnitude is small in our LPM model the result indicates a significant result.

But what does this mean for teams that are buying players for the next season? Well, a typical team would have a starting pitcher who pitches most of the game and relief and a closer pitcher after the pitcher reaches a certain pitch count or inning. What our result suggests when picking out the bullpen the overall ERA would be more important than having a couple of good starters and bad relief and closers. The game is won not by starters but by the overall performance of the bullpen. Offensively the average runs scored statistic would be the most important, but realistically the focus would be on the pitcher rotation as it's possible ton win with a low run scored as long as the ERA is low.

7 Conclusion

In this study, we attempt to determine the key statistics that predict a Major League Baseball (MLB) team's chances of winning their division and by extension a post-season spot. By using logistic regression models and analyzing data spanning 22 seasons(1998-2019) from Lahman's Baseball Database, we attempted to focus on the ERA and runs scored as the most significant indicators.

Our findings suggest ERA and runs scored are important variables in predicting whether a team wins the division. Teams with lower ERAs and higher runs scored have a significantly higher probability of winning their division. These results are consistent across multiple models and remain robust even when testing for potential multicollinearity, particularly when modeling one offensive and one defensive statistic in our model.

The impact of our study is on how teams should be team building and player scouting. Teams should focus on building a well-rounded bullpen to maintain a low ERA across the entire pitching staff, rather than relying on a few star starter pitchers. Offensively, while runs scored are important, it should be on a balanced approach that looks to consistent scoring opportunities throughout the season as the low ERA and average runs scored can lead to a division win.

In conclusion, our research supports the importance of ERA on the probability of winning if our statement can be taken as casual. By focusing on these key indicators, small-market MLB teams can make more informed decisions during free agency and player acquisitions, leading to better overall team performance and increased chances of post-season success with a smaller budget. Future studies could expand on this by adding additional variables and exploring their impact on other post-season outcomes and possibly creating more robust statistics that combine other statistics like OERA to test for.

8 Appendix

Figure 1: Different Types of Running Variables

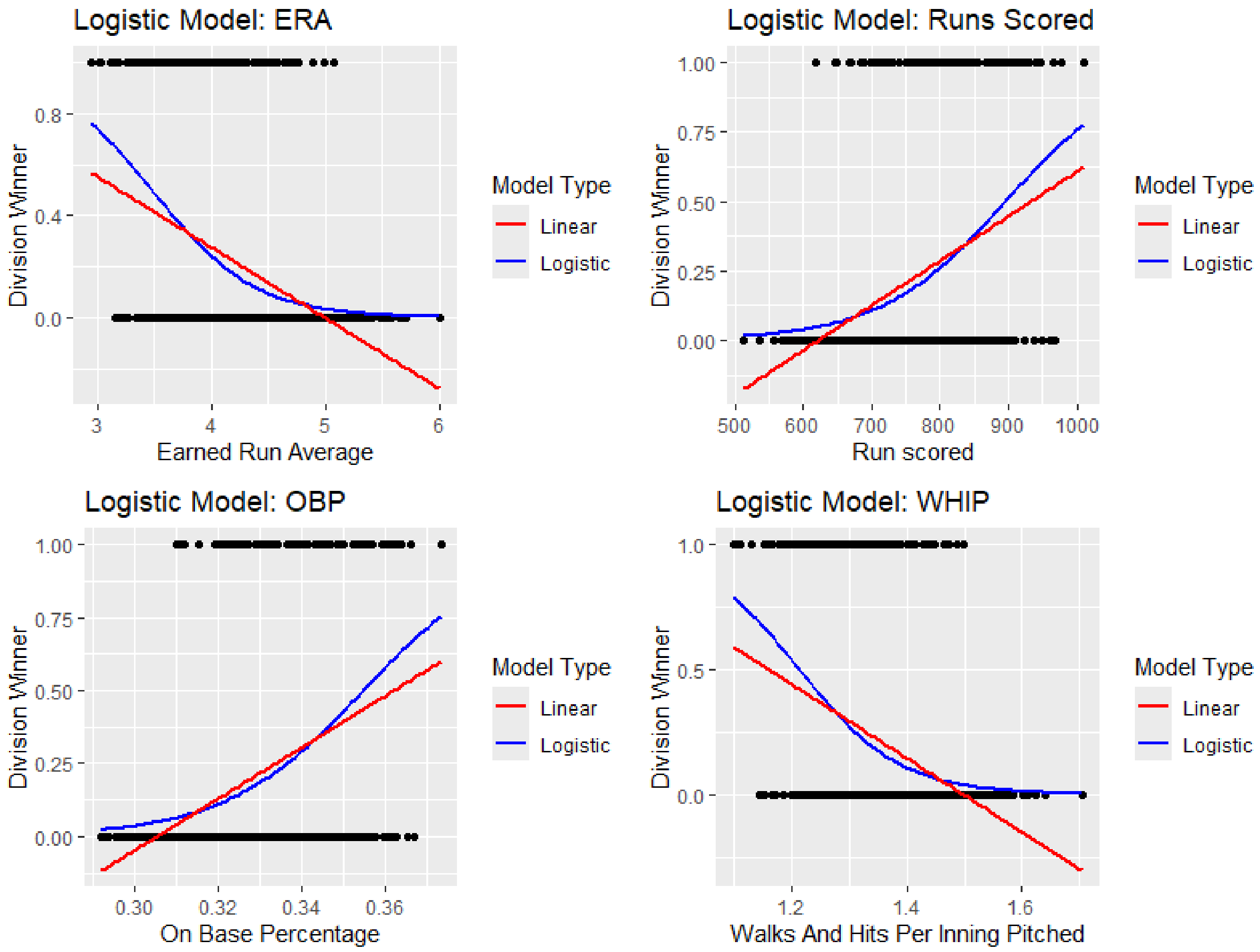

Table 4

	GVIF	Df	GVIF^(1/(2*Df))
runs_scored	12.206	1	3.494
obp	8.972	1	2.995
homeruns	4.520	1	2.126
earned_run_average	9.122	1	3.020
whip	8.259	1	2.874
factor(year)	21.941	21	1.076
factor(team_name)	21.415	33	1.048

Table 5

	GVIF	Df	GVIF^(1/(2*Df))
runs_scored	11.611	1	3.407
obp	8.280	1	2.877
homeruns	3.982	1	1.996
factor(year)	6.654	21	1.046
factor(team_name)	4.646	33	1.024

Table 6

	GVIF	Df	GVIF^(1/(2*Df))
earned_run_average	7.470	1	2.733
whip	7.595	1	2.756
factor(year)	4.560	21	1.037
factor(team_name)	3.133	33	1.017

Table 7

	GVIF	Df	GVIF^(1/(2*Df))
whip	2.288	1	1.513
factor(year)	2.921	21	1.026
factor(team_name)	2.130	33	1.012

Table 8

	GVIF	Df	GVIF^(1/(2*Df))
earned_run_average	2.293	1	1.514
factor(year)	2.970	21	1.026
factor(team_name)	2.477	33	1.014

Table 9

	GVIF	Df	GVIF^(1/(2*Df))
runs_scored	2.926	1	1.711
factor(year)	3.394	21	1.030
factor(team_name)	2.688	33	1.015

Table 10

	GVIF	Df	GVIF^(1/(2*Df))
obp	2.699	1	1.643
factor(year)	3.305	21	1.029
factor(team_name)	2.101	33	1.011

Table 11

	GVIF	Df	GVIF^(1/(2*Df))
runs_scored	3.763	1	1.940
earned_run_average	2.993	1	1.730
factor(year)	7.448	21	1.049
factor(team_name)	7.148	33	1.030

References

Albert, J. (2010). Sabermetrics: The past, the present, and the future.

Cover, T. M. and Keilers, C. W. (1977). An offensive earned-run average for baseball. *Operations Research*, 25(5):729–740.

Houser, A. (2005). Which baseball statistic is the most important when determining team s. *The Park Place Economist*, 13. Accessed: 2024-04-28.

Lindsey, G. R. (1959). Statistical data useful for the operation of a baseball team. *Operations Research*, 7(2):197–207.