Thomas Farley
Dr. Tammy Leonard
Econometrics
3 May 2021

The Team Pitching Statistics That Seemingly Affect Win Count

I. Introduction

In the sport of baseball, especially the major leagues, good pitching on an individual and team level leads to more player and team wins respectively. Baseball Reference[1], the source of my data, shows statistically that this is indeed true. However, since baseball is a very dynamic and multi-faceted sport, it is unclear which statistics are the best in determining the value of a team. That leads me to question: what individual pitching statistic contributes most to a team's wins? In this study, I wish to show which team pitching statistic is the most valuable for general managers when they are aiming for a high win-count season.

II. Theoretical Basis

Teams who make the playoffs have some of the best pitching out of the entire league, as well as the World Series Champions the year that they win. Pitching is just one of the many dynamics of the sport; it is a realm of its own and thus requires much analysis. Since it is extremely difficult for a player singlehandedly to impact the win count of his team (an analytical fact that is regularly known to this day by sabermetricians in MLB), there is more value in analyzing a team's pitching statistics rather than an individual player's. For example, the New York Mets had the MLB's best pitcher in 2019, Jacob deGrom, yet they had a losing record and did not make the playoffs. Their team had some of the worst team pitching. The scope of the project is not to determine who the best team is, but rather the best pitching statistic; we already know that good pitching means more wins. If a general manager needs to sign pitchers to a contract, then what should he look for in all the free agent pitchers? The aim of this project is to find out which statistic, whether it is ERA+, FIP, or WHIP (all commonly looked at stats), is all around the best in determining team pitching value.

III. Data

The following data comes from a reputable baseball statistic website that I have used in the past called Baseball Reference. Each team value shown below is from the year 2019, since that is the last full season played. The main dependent variable that I plan to analyze is win count per team. I have chosen relatively popular stats among baseball nuts like me such as FIP, which is fielding independent pitching, and HR9, home runs allowed per nine innings as stated earlier. All of these are independent variables and they are all continuous.

---

[1] "2019 Major League Baseball Season Summary." *Baseball Reference*, www.baseball-reference.com/leagues/MLB/2019.shtml.

Table 1. Description of Each Independent Variable

RA/G	Runs allowed per game
ERA+	ERA compared to rest of the league with 100 being league average
FIP	Fielding independent pitching
WHIP	Walks + hits per inning
H9	Hits allowed per nine innings
HR9	Home runs allowed per nine innings
BB9	Walks allowed per nine innings
SO9	Strikeouts per nine innings

Table 2. Summary Statistics of Continuous Variables

	W	**RA/G**	**ERA+**	**FIP**	**WHIP**	**H9**	**HR9**	**BB9**	**SO9**
Standard Dev	15.90	0.60	12.00	0.45	0.10	0.65	0.18	0.36	0.69
Average	80.97	4.83	102.20	4.51	1.33	8.72	1.41	3.30	8.88
Minimum	47	3.78	83	3.65	1.102	7.4	1.1	2.4	7.7
Maximum	107	6.06	127	5.56	1.494	9.8	1.9	3.8	10.3

IV. Econometric Data

A linear regression model was estimated to investigate the impact of team pitching statistics on team win count. The basic model is expressed as Equation 1 below. I also included a log-linear model (Equation 2) and a log-log model (Equation 3):

$$W_i = \beta_0 + \beta_1 era_i + \beta_2 fip_i + \beta_3 h_i + \beta_4 hr_i + \beta_5 walk_i + \beta_6 so_i + \epsilon_i \quad (1)$$

$$ln(W_i) = \beta_0 + \beta_1 era_i + \beta_2 fip_i + \beta_3 h_i + \beta_4 hr_i + \beta_5 walk_i + \beta_6 so_i + \epsilon_i \quad (2)$$

$$ln(W_i) = \beta_0 + \beta_1 ln(era_i) + \beta_2 ln(fip_i) + \beta_3 ln(h_i) + \beta_4 ln(hr_i) + \beta_5 ln(walk_i) + \beta_6 ln(so_i) + \epsilon_i \quad (3)$$

As noted previously, the dependent variable is win count. The covariates are ERA+ (*era*), fielding independent of pitching (*fip*), hits allowed per nine innings (*h*), home runs allowed per nine innings (*hr*), walks allowed per nine innings (*walk*), and strikeouts per nine innings (*so*). Each variable is considered an independent variable of interest, since discovering which pitching stat contributes the most to wins is the goal of this project. A log-log model was estimated because I presumed that since ERA+ is a mean compared to other parts of the league, then there would be a normal distribution. Since there is a normal distribution, then running a log model

should give us better correlation if a normal distribution follows a bell-curve equation. I ended up not including the stats WHIP and RA/G because I feel as though they would be confounders and lead to bias in the results. WHIP is walks plus hits per inning pitched and since I am already including hits per nine and walks per nine, I would technically be including the same thing twice. This is the same for RA/G, since ERA+ is a league comparison of ERA, which is earned runs per nine innings.

The natural log of each variable, both dependent and independent was calculated in a separate column in Excel. This is so Excel can accurately apply the variables to Model 2 and Model 3, which include the log of one and all variables respectively.

V. Results

All three models were estimated. Table 3 represents the estimation results for these three models. Model 1 includes the results from Eq. 1, as it is a simple multiple linear regression model. Model 2 includes the results from Eq. 2, which is the log linear model. Model 3 includes the results from Eq. 3, which is the log-log regression model. All models include each variable of interest.

The coefficient estimates for the covariates are relatively similar across the three models, though the standard error is not quite. Wins seem to be associated with a lower number of team walks, hits, and home runs per nine innings as well as FIP while they also seem to be associated with a higher number of team ERA+ and strikeouts per nine innings. The coefficient estimates for fip_i, h_i, hr_i, $walk_i$, and so_i were determined to be not statistically significant for all three models. The adjusted R-squared for each model number respectively were 0.643, 0.591, and 0.607.

The only statistically significant statistic is ERA+ in any of the three models, meaning the coefficient on era_i is statistically significant. In Model 1, a 5.00% increase in ERA+ on average leads to a 4.55% increase in wins. Model 2 shows that a 5.00% increase in ERA+ leads to a 4.73% increase in wins on average. Lastly, Model 3 shows that a 5.00% increase in ERA+ will get one a 5.25% increase in wins on average. ERA+ has a steady standard error compared to the other variables.

VI. Discussion

Since the only statistically significant statistic is ERA+, then ERA+ could possibly lead to multicollinearity issues. However, new regressions were calculated using the same three models listed above (Model 1, Model 2, and Model 3) but without the era_i variable. Without ERA+ as an included variable, the coefficient estimates for fip_i, h_i, hr_i, $walk_i$, and so_i were even still determined to be not statistically significant for all three models. Since ERA+ was originally believed to be causing multicollinearity, I now no longer believe that it is since without it included, the significance and the probability of the other variables barely change *with* ERA+ included as a variable.

Though teams that typically allow many home runs, hits, or walks tend to have higher team ERAs than teams that do not, that does not mean that a team will accumulate a higher win count in the long run. This is also true for strikeouts but positively correlated; teams that accumulate more strikeouts have lower ERAs, but not necessarily a higher win count. Though ERA+ is not our dependent variable, it does seem to already have a possible correlation with the other covariates, but as stated earlier, that does not seem to be the case. According to the statistical significance of the coefficient of era_i, having a higher ERA+, which would imply allowing fewer runs than the rest of the league, correlates to having more wins. It is true that you get a win if you outscore the other team in a single game, but ERA+ demonstrates how well you limit the other team's run count in a game, not how much you outscore the other team. A team may hold most opposing teams to few runs on average, but may be held to few runs themselves and not gain many wins. This study, however, shows that having a high ERA+ correlates to having more wins on average. The other variables are not that significant. Striking out an opponent a lot will not lead to more wins. Giving up a lot of home runs will not necessarily mean that a team will lose more.

It is interesting to note the lack of statistical significance on the variable FIP. Fielding independent pitching has recently been a highly touted stat by sabermetricians for it measures the events that the pitcher has complete control over.[2] These events are *only* strikeouts, home runs, intentional and unintentional walks, and hit-by-pitches. A pitcher with a low FIP is good because he has great control over his command and is more likely to pitch to a desired outcome. One would think this stat would be statistically significant, but it is very interestingly not. This is likely because of those events that the pitcher does not have complete control over, such as singles, doubles, triples, inside-the-park home runs, fielding errors. This could indicate that most pitchers, if not all, rely heavily on the fielders backing them up behind them. Imagine a baseball team where it is just the pitcher. That seems silly but that is part of the explanation for why FIP is likely not that significant as everyone thinks. FIP is better used to compare individual players on the same team or among other teams.

This study is not without limitations. There are many other raw statistics that could be gathered, such as average barrel percentage or average exit velocity, but it would be extremely difficult to compile them since they belong to other databases such as Baseball Savant and PITCHfx. Also, since baseball is a very unpredictable sport, it could be difficult to accurately explore interaction variables. A further study of offensive statistics must be considered for determining correlation to win count. It would be interesting to do a study that compares both team offensive statistics to pitching statistics and how they might contribute.

Overall, a higher ERA+ will on average lead to a higher win count.

VII. Conclusion

[2] "Fielding Independent Pitching (FIP): Glossary." MLB.com, www.mlb.com/glossary/advanced-stats/fielding-independent-pitching.

In conclusion, a higher team ERA+ will on average lead to a higher win count for a team. When one hears the phrase, “defense wins championships,” this study only shows that that phrase is partly true, at least in regards to team ERA+ for team pitching.

Table 3. Estimation Results

	Model 1	Model 2	Model 3
ERA+	0.7300**	0.0089*	
	(0.289)	(0.004)	
FIP	-8.035	-0.152	
	(47.1)	(.661)	
H9	-0.305	-0.012	
	(4.938)	(0.069)	
HR9	0.774	0.025	
	(67.09)	(0.942)	
BB9	-5.267	-0.023	
	(19.151)	(0.269)	
SO9	0.101	-0.011	
	(11.573)	(0.162)	
ln(ERA+)			1.050**
			(0.416)
ln(FIP)			0.808
			(2.609)
ln(H9)			0.033
			(0.585)
ln(HR9)			-0.537
			(1.127)
ln(BB9)			-0.155
			(0.248)
ln(SO9)			0.626
			(1.29)
N	30	30	30
Adjusted R Squared	0.643	0.591	0.607

$\alpha = 0.05 \rightarrow$ *, $\alpha = 0.025 \rightarrow$ **, $\alpha = 0.01 \rightarrow$ ***