Which Statistics in Baseball Are Most Important for Winning?

Clark Ballou-Crawford, Fort Hays State University

Abstract

I investigate which professional baseball team stats have the greatest correlation with regular season wins over the previous six full MLB seasons. This analysis is performed by gathering team-by-team season statistics to be utilized as explanatory variables (a full list is included below) and regress these against the dependent variable for each team, regular season wins. Through this analysis it is determined that a set of five of these variables can create a full regression model with a coefficient of determination of .882, implying that 88.2% of the variability in regular season win totals can be explained using these five team statistics. Power hitting has become a point of emphasis for team builders in recent years and this trend is supported by this analysis.

Introduction

Professional baseball owners and managers are always searching for a competitive advantage with which to arm their teams. The main way they pursue these advantages is by using statistics. In this age of information, every action by players on the baseball diamond is somehow quantified and then ranked and evaluated, giving team planners insight as to how their teams are performing. The goal of this analysis is ultimately to win more games. This study attempts to provide insight into which or what combination of these statistics are most important for regular season wins. This is done by regressing a dependent variable, regular season wins, on six numerical variables and one categorical variable that are described below. Summary statistics are provided in the next table.

Variable	Type	Description
BA	Numerical	Team batting average
OPS+	Numerical	Team on-base % + slugging %, normalized
HR	Numerical	Team home runs hit
WHIP	Numerical	Walks + hits allowed/inning pitched
HRA	Numerical	Home runs against
ERA+	Numerical	Earned run average/9 innings, normalized
NL	Categorical	1 if team is NL, 0 if AL

Summary Statistics

	General Info			*Quantiles*				
Variable	**Obs.**	**Mean**	**S.D.**	**Min**	**0.25**	**Mdn**	**0.75**	**Max**
WINS	180	81	14	47	71	81	91	111
HR	180	196	37	110	168	198	221	307
HRA	180	196	29	132	176	192	214	305
OPS+	180	97	9	77	92	97	104	123
ERA+	180	102	13	77	93	101	110	150
BA	180	0.25	0.01	0.22	0.24	0.25	0.26	0.28
WHIP	180	1.31	0.1	1.05	1.24	1.31	1.38	1.5

VARIABLES	(1) WINS	(2) WINS	(3) WINS	(4) WINS	(5) WINS
HR	0.199***	0.225***	0.195***	0.112***	0.108***
	(0.0233)	(0.0169)	(0.0161)	(0.0142)	(0.0118)
HRA		-0.272***	-0.267***	-0.126***	-0.0934***
		(0.0211)	(0.0192)	(0.0191)	(0.0163)
BA			323.5***	237.4***	334.9***
			(51.88)	(40.16)	(35.07)
ERA				0.552***	0.188***
				(0.0485)	(0.0573)
WHIP					-61.98***
					(6.927)
Constant	42.07***	90.22***	14.44	-32.08***	56.74***
	(4.638)	(5.011)	(12.97)	(10.68)	(13.31)
Observations	180	180	180	180	180
R-squared	0.290	0.634	0.700	0.828	0.882

Standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1

Methodology & Model

Data for this study comes from baseball-reference.com, an online database of baseball statistics and anomalies. The sample used in this study is comprised of team-by-team data from each of the previous six full MLB seasons (2020 was a shortened season and its data was not considered), equaling 180 total samples in all. Analysis is performed through regression and estimated using the ordinary least squares method.

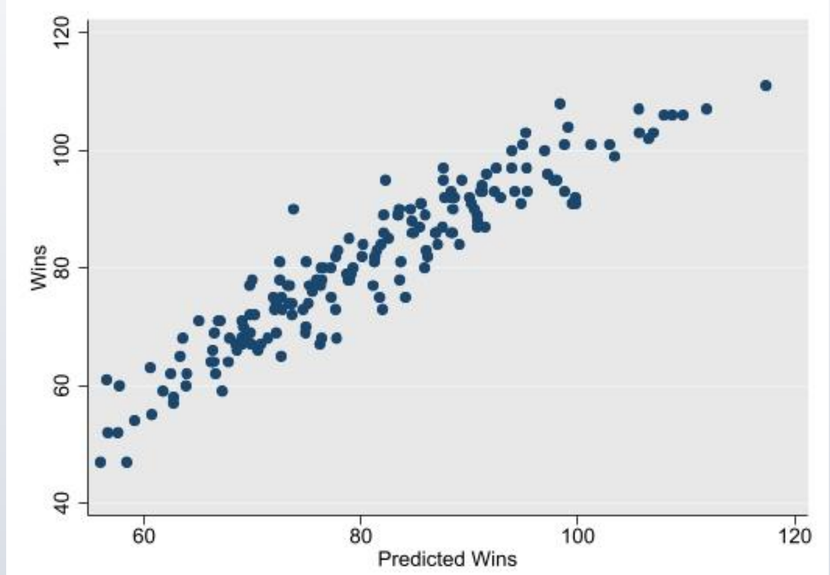
Selected Results

Prior to creation of the final model a couple variables that were considered had to be eliminated from consideration. The categorical variable NL was found to be insignificant at all conventional levels of alpha, implying that there is no statistical difference in wins for a team in the National League compared to a team in the American League. The other eliminated variable, OPS+, was decided to have too much multicollinearity with other variables, namely BA and HR, that resulted in erratic p-values and high mean standard errors when examining different models and was thus excluded.

Summary statistics indicate that there is a considerable amount of variation in season home run totals for teams, both offensively and defensively (HR and HRA, respectively). Through regression analysis it is determined that both variables have a strong correlation coefficient with the dependent variable, wins, which implies a strong relationship between the variables. A model with only these two variables regressed against wins has a coefficient of determination of .634, meaning 63.4% of the variability in regular season win totals is explained by differences in these two statistics. This result supports recent trends in the sport towards emphasizing power hitting by batters and limiting powerful hits off pitchers.

The final model includes the variables HR, HRA, WHIP, ERA+ and BA. All included variables are statistically significant at all conventional levels of alpha and the final model minimized the standard error among all tested models. The final model accounts for 88.2% of the variance in wins, and can be represented by the following equation:

$$\boldsymbol{WINS} = \mathbf{56.74} + \mathbf{.108}\,(\boldsymbol{HR}) - \mathbf{.0934}\,(\boldsymbol{HRA}) + \mathbf{334.9}\,(\boldsymbol{BA}) + \mathbf{.188}\,(\boldsymbol{ERA}+) - \mathbf{61.98}\,(\boldsymbol{WHIP})$$



Conclusion

Results indicate that MLB teams would be wise to emphasize power hitting on offense and limiting the number of powerful hits their pitchers surrender. This conclusion is slightly different than similar studies in recent years that concluded that power hitting wasn't strongly correlated with winning but does coincide with recent industry trends toward more power hitting and less volume hitting. This, though, comes with the caveat that other measures of team performance are important as well and shouldn't be neglected in search of power. This model concludes that a team's ERA+, BA, and WHIP are also statistically significant in the search for wins, and this is only of the variables that were examined. Further study is required to determine which additional statistics would further improve the model, and by how much.

Selected References

MLB stats, scores, History, & Records. Baseball. (n.d.). Retrieved December 9, 2022, from https://www.baseball-reference.com/

Sult, S. (2021, May 9). *Common MLB statistics: Which stats determine a team's win percentage?* Medium. Retrieved December 9, 2022, from https://sarahesult.medium.com/common-mlb-statistics-which-stats-determine-a-teams-win-percentage-a6e0a83aa07c

Williams, C. (2019, June 17). *How important are home runs in a power hitting period?* Samford University. Retrieved December 9, 2022, from https://www.samford.edu/sports-analytics/fans/2019/How-Important-Are-Home-Runs-in-a-Power-Hitting-Period