Minnesota Intercollegiate Athletic Conference (MIAC) Softball: An Evaluation of Statistics for Playoff Berths and Winning Percentage.

Abstract

Although the current use of statistics in baseball is extensive, a similar approach in softball is much less developed. In this paper, selected softball statistics (representing various offensive, defensive, and pitching measures) from the Minnesota Intercollegiate Athletic Conference (MIAC, NCAA Division III) are analyzed to determine the relationship between these statistics and team success over the course of a season. Through the use of two-sample t-tests and multiple regression, we find that (a) seven key statistics separate MIAC softball teams that qualify for playoffs from those that do not; and (b) three key statistics are especially important in predicting team winning percentage. Moreover, the three key statistics from (b) represent all three softball skill areas: batting, fielding, and pitching; suggesting that good teams cannot overlook any area. We indicate several directions for future research, which include the use of logistic regression and the incorporation of more advanced softball statistics.

Introduction

The use of statistics to help determine the possibility of teams winning athletic contests is not a new idea. The multi-million dollar business of Major League Baseball has been using players' individual statistics to determine a player's worth for decades. In order for a team to win, however, they simply must score more runs than their opponent. So, although there are many different recorded statistics in baseball, some are more useful than others for evaluating overall team performance [1,2]. While the current use of statistics in baseball is extensive, a similar use of statistics in softball is less often employed. Although statistics are used, the field of sabermetrics is much more developed for baseball than for softball. In particular, there is great opportunity to explore statistics in NCAA Division III softball, in the Minnesota Intercollegiate Athletic Conference (MIAC). Based on various studies [1,2,3], we selected several softball statistics to investigate our primary research questions of: what offensive, defensive, and pitching statistics are most important for MIAC softball teams for (1) securing a MIAC playoff berth; and (2) predicting team winning percentage?

Materials and Methods

Our data was obtained from the MIAC softball archives.[1] Since MIAC playoff data exists for the past eleven seasons (2008 season - 2018 season), it is data from within this time frame that we incorporated into our data set.[2] Data for offensive, defensive, and pitching statistics were scraped from the web using R (either directly, or created using data manipulation) and organized into CSV files according to statistic type.

Given our primary research questions (as above), we explored two response variables: (1) whether a team qualifies for playoffs (categorical - yes/no); and (2) team winning percentage (numerical; winpct). We also examined seven explanatory variables: team on-base percentage (OBP), runs scored per game (RPG), walk-to-strikeout ratio (BBK), and slugging percentage (SLG) (offensive measures); team fielding percentage (FPCT) (defensive measure); and pitching staff walks plus hits per inning pitched (WHIP) and earned run average (ERA) (pitching measures). (See Appendix: Table 1, for a full list and accompanying descriptions of the variables of interest.)

We used two main statistical inference methods to quantify the association between our outcome and predictor variables. We used two-sample t-tests (as well as 95% t-confidence intervals (t-CI's)) to compare the seven explanatory variables between teams that qualified for playoffs and teams that did not. We also used a backward-elimination process to obtain a multiple regression model for team winning percentage (winpct), which accounts for the three predictors of RPG, FPCT, and ERA.

Results

In our sample, the average number of runs scored per game for playoff teams was 6.13 with a range of 5.13, compared to a mean number of runs scored per game for non-playoff teams of 3.64 with a range of 4.78. The mean fielding percentage for playoff teams was .962 with a range of .047, while the mean fielding percentage for non-playoff teams was .945 with a range of .073.

---

[1] https://www.miacathletics.com/sports/sball/archive.

[2] For the purposes of this project, we have chosen to not use the data from the 2012-2013 season. Due to weather, for the 2012-2013 season only, the MIAC adopted a 12-team format for playoffs, as opposed to the typical 4-team format.

The mean ERA for playoff teams was 2.36 with a range of 3.64, as compared to a mean ERA for non-playoff teams of 4.59 with a range of 9.94. (See Appendix: Table 2, for a summary of the results from our exploratory data analysis for our key explanatory variables.)

We performed seven two-sample t-tests (one for each key explanatory variable) to compare softball statistics between playoff and non-playoff teams. One particularly interesting finding was that the t-test for RPG revealed that there is statistically significant evidence (T = -11.486, p < 2.2e-16, 81.06 df) that, on average, playoff teams score more runs per game than non-playoff teams. We are 95% confident that the true mean number of runs scored per game is between 2.06 and 2.92 runs per game higher for playoff than non-playoff teams (see Figure 1). Results from the two-sample t-tests for the other key explanatory variables lead to similar interpretations (see Appendix: Table 3). Overall, from our seven two-sample t-tests, we found that there is statistically significant evidence that, on average, the OBP, RPG, BBK, SLG, and FPCT are each higher for playoff than non-playoff teams; whereas the true mean WHIP and ERA are each lower for playoff than non-playoff teams.

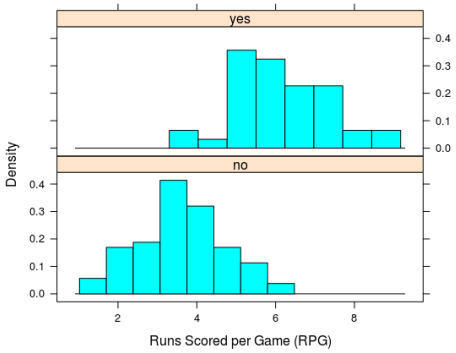

Figure 1. Runs Scored per Game for Playoff (yes) and Non-Playoff (no) MIAC Softball Teams.

We obtained a final multiple regression model as given below:

$$\hat{\text{winpct}} = -0.691356 + 0.083200*(\text{RPG}) + 1.117488*(\text{FPCT}) - 0.066251*(\text{ERA}).$$

This model accounts for three explanatory variables: RPG, FPCT, and ERA (see Appendix: Table 4). Overall, 91.76% of variability in winning percentage is explained by this model. This model indicates that a one run increase in RPG is associated with a 0.083 increase in winning percentage, after accounting for FPCT and ERA. Similarly, a .1 increase in FPCT is associated with a .112 increase in winning percentage, after accounting for RPG and ERA. Also, a one run increase in ERA is associated with a 0.066 decrease in winning percentage, after accounting for RPG and FPCT. From our analysis, there is statistically significant evidence that better statistics for all of RPG (T = 15.810, p-value < 2e-16), FPCT (T = 2.008, p-value = 0.0469), and ERA (T = -11.294, p-value < 2e-16) are associated with an increase in winning percentage, after accounting for the other two explanatory variables. That is, one each of offensive, defensive, and pitching measures is statistically significant after accounting for the others. (See Appendix: Figure 4, for scatter plots of the variables in the final multiple regression model.)

We used the final multiple regression model to make predictions for the 2019 MIAC softball season. The full results of these predictions, along with the actual results from the 2019 MIAC softball season are given in the Appendix: Table 5. One interesting finding from a comparison of the actual and predicted results involves St. Olaf College (St. Olaf) and the College of St. Benedict (St. Ben's). St. Olaf and St. Ben's had actual winning percentages of .727 and .636, respectively, finishing tied for 2nd and 5th in the conference, again respectively. In contrast, our model predicted winning percentages of .514 and .671 for St. Olaf and St. Ben's, corresponding to 6th and 3rd place finishes in the conference. Notably, from the resulting conference standings, our model predicted that St. Olaf would not qualify for playoffs while St. Ben's would, which is in direct opposition to the actual outcome. So, while the predictions of our model agree in many cases with the actual 2019 overall trends, certain discrepancies such as the St. Olaf/St. Ben's prediction suggest that our model also has the ability to identify teams that overperform or underperform statistical expectations.

Discussion

Several results from our analysis provided insight on key statistics that separate playoff and non-playoff softball teams in the MIAC, and indications of which may also be most useful in predicting team winning percentage, thereby answering our primary research questions. Our results suggest that, in MIAC softball, the true mean OBP, RPG, BBK, SLG, and FPCT are higher for playoff than non-playoff teams; and that the true mean WHIP and ERA are lower for playoff than non-playoff teams. Our analysis suggests that the three most important of our key variables in predicting winning percentage for MIAC softball teams are RPG, FPCT, and ERA (one offensive, one defensive, and one pitching measure). Moreover, the fact that these three predictive variables span all three softball skill areas (offense/batting, defense/fielding, and pitching) is especially notable. This suggests that good teams cannot overlook any of these areas; that a strength in one or even two of the skill areas may carry a team only so far.

Our findings support several results from the literature, especially the "runs-created" approach to offensive statistics. This approach focuses on statistics that "create runs" for a team, rather than strictly examining individual batters' success at the plate [3]. For instance, our final multiple regression model suggests that, as an offensive measure, RPG is more important (than OBP) as a predictor of team winning percentage. This result also aligns with the finding that batting average (BA) is not necessarily indicative of team wins [1], since both OBP and BA are similarly concerned with a player's ability to get on base. Moreover, two of the three softball statistics from our final multiple regression model were more focused on runs (RPG and ERA) than their offensive or pitching counterparts (OBP, BBK, SLG, and WHIP).

Based on our data collection methods, it is reasonable to generalize our findings to MIAC softball. Our study, however, has some limitations. It may be that, for women's softball, the MIAC is similar to other NCAA Division III conferences. Yet other factors such as weather and field conditions may render the generalizability of our findings to all of Division III softball questionable. Additionally, as an observational study, we cannot conclude causation. Still, given our methods and the context of other literature, our study suggests several paths for future research. First, logistic regression could provide an effective means to investigate how well particular statistics may or may not be predictive of whether a team qualifies for playoffs, as well as to examine the outcomes of head-to-head matchups. A number of other factors could also be incorporated into future study. Such factors include: potential confounders (e.g., BA w/ runners in scoring position, baserunning), and seasonal and situational considerations (e.g., results from previous seasons, runners left on base, hitting streaks) [2,4]. Further research could also explore the use of more advanced statistics, as several pieces of literature have suggested better ways to measure defensive effectiveness and forecast pitcher performance (than FPCT and ERA) [3].

References

[1] Bennett, J., & Flueck, J. (1983). An Evaluation of Major League Baseball Offensive Performance Models. The American Statistician, 37(1), 76-82. doi:10.2307/2685850.

[2] Otten, & Barrett. (2013). Pitching and clutch hitting in Major League Baseball: What 109 years of statistics reveal. Psychology of Sport & Exercise, 14(4), 531-537.

[3] Winston, W. (2009). Mathletics: How Gamblers, Managers, and Sports Enthusiasts Use Mathematics in Baseball, Basketball, and Football. Princeton University Press.

[4] Albert, J. (1994). Exploring Baseball Hitting Data: What About Those Breakdown Statistics? Journal of the American Statistical Association, 89(427), 1066-1074. doi:10.2307/2290936.

Appendix

Variable Name	Variable Role	Description
playoffs	Response	yes/no; based on whether a team qualifies for MIAC playoffs
winpct	Response	team winning percentage
OBP	Explanatory	on-base percentage (walks plus hits plus hit-by-pitches divided by plate appearances)
RPG	Explanatory	runs scored per game
BBK	Explanatory	walk-to-strikeout ratio
SLG	Explanatory	slugging percentage; gives more weight to extra base hits ((1B + 2x2B + 3x3B + 4x4B)/AB)
FPCT	Explanatory	fielding percentage (putouts plus attempts divided by total chances)
WHIP	Explanatory	walks plus hits per inning pitched
ERA	Explanatory	earned run average (number of earned runs given up per seven innings pitched)

Table 1: Variables (all statistics from MIAC conference play).

	Mean Value		Range	
Statistic	**playoff teams**	**non-playoff teams**	**playoff teams**	**non-playoff teams**
OBP	.393	.323	.14	.16
RPG	6.13	3.64	5.13	4.78
BBK	.835	.457	1.24	.595
SLG	.463	.358	.296	.243
FPCT	.962	.945	.047	.073
WHIP	1.31	1.72	1.26	1.81
ERA	2.36	4.59	3.64	9.94

Table 2. Summary of Exploratory Data Analysis (EDA) for Key Explanatory Variables.

Statistic	T test statistic (T)	p-value (p)	degrees of freedom (df)	95% t-confidence interval (t-CI)
OBP	-10.752	< 2.2e-16	92.739	(0.057, 0.083)
RPG	-11.486	< 2.2e-16	81.06	(2.06, 2.92)
BBK	-7.0382	5.754e-09	49.057	(0.27, 0.49)
SLG	-10.058	1.493e-15	74.968	(.098, 0.146)
FPCT	-6.9763	3.003e-10	103.02	(0.013, 0.023)
WHIP	9.6646	5.679e-16	99.412	(0.400, 0.607)
ERA	9.7049	< 2.2e-16	116.32	(1.77, 2.68)

Table 3. Summary of Two-Sample T-Tests for Key Explanatory Variables.

Statistic	Slope	95% CI for Slope	T test statistic (T)	P-value (p)
RPG	0.083200 (0.083)	(0.073, 0.094)	15.810	< 2e-16
FPCT	1.117488 (1.117)	(0.015, 2.220)	2.008	0.0469
ERA	0.066251 (0.066)	(0.055, 0.078)	-11.294	< 2e-16

Table 4. Summary of Explanatory Variables in the Final Multiple Regression Model for Winning Percentage.

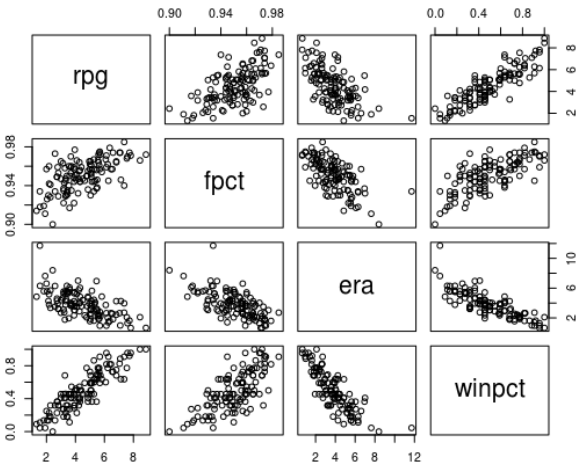

Figure 2. Scatter Plots for Winning Percentage and Explanatory Variables from the Final Multiple Regression Model.

	Actual			Predicted		
Team	**Standings**	**Winning %**	**Playoffs**	**Standings**	**Winning % / 95% PI**	**Playoffs**
St. Thomas	1	.864	yes	1	.819 / (0.679, 0.958)	yes
Hamline	T2 (2)	.727	yes	4	.605 / (0.466, 0.745)	yes
St. Kate's	T2 (3)	.727	yes	2	.696 / (0.556, 0.835)	yes
St. Olaf	T2 (4)	.727	yes	6	.514 / (0.374, 0.654)	no
St. Ben's	5	.636	no	3	.671 / (0.532, 0.811)	yes
Bethel	T6 (6)	.455	no	5	.594 / (0.456, 0.732)	no
Carleton	T6 (7)	.455	no	7	.495 / (0.357, 0.633)	no
Gustavus	T8 (8)	.364	no	8	.437 / (0.298, 0.575)	no
Macalester	T8 (9)	.364	no	9	.343 / (0.203, 0.482)	no
Augsburg	10	.318	no	11	.322 / (0.183, 0.461)	no
St. Mary's	11	.273	no	10	.332 / (0.193, 0.470)	no
Concordia	12	.091	no	12	.028 / (-0.112, 0.169)	no

Table 5: Actual and Predicted MIAC Softball Results for the 2019 Season.