What Makes the Minnesota Twins a Successful Baseball Organization?

Keegan Murray, Walter Fenske

Introduction

Something incredible happened recently: the Twins won their first playoff game since 2004.[1] How does a team break a 19 year playoff drought? This immediately intrigued us as to how their 2023 performance was so different. We aim to answer the question: What makes the Minnesota Twins Baseball Organization successful? Alongside this, we want to know if 2023 was an outlier year for the Twins. Our results can inform us of which baseball statistics are most effective at predicting win percentage and if the Twins had an exceptional season in 2023.[6]

Dataset and Exploratory Data Analysis

Data and Variables

We are looking at Minnesota Twins stats from 1992-present, since the last time the Twins won the world series was 1991. It also does not make full sense to look at data from the entire existence of the Twins as baseball is played differently today than it was 75+ years *ago*.[3] This data was collected by Major League Baseball and consists of *batting*[4] and *pitching*[5] average statistics per year in the selected years. Some of these statistics include wins, losses, hits for, hits against, and many other predictors of win percentage. We calculate a new variable called $WinPct$, or win percentage, by using the aforementioned statistics. Win percentage will be the response variable while the remaining 32 variables are potential explanatory variables.

We decided to exclude the year 2020 since this was the year COVID-19 ramped up, meaning only 60 games were played rather than the usual 162. We also chose to exclude 1994 and 1995 since only 113 and 144 games were played respectively due to a Major League Baseball *strike*.[2]

Fig. 1 explains the names of the variables which ended up in our final models. The only variables that need further explanation are SLG (slugging) and SV (saves). Many other variables are self explanatory. With that in mind, a team receives a save if one of the following is true:

- Enters the game with a lead of no more than three runs and pitches at least one inning.
- Enters the game with the tying run in the on-deck circle, at the plate or on the bases. This essentially means that there are one or more runners one base. If the batter is succesful, the score will be tied.
- Pitches at least one inning.

Another complex stat is Slugging Percentage (SLG). Slugging percentage is calculated by the total number of bases hit divided by the total number of at bats. This means that if a player has a 1.0 SLG, they average one base recorded per at bat on average. This would be very high as the twins average 0.427 SLG, meaning that that on average they get 0.427 bases per at bat.

Exploratory Data Analysis

We were chiefly interested in exploring the relationship between win percentage and any numeric variables that appeared to have a strong ($r > 0.5$) correlation coefficient. To do this, we created two separate correlation matrices, one for batting statistics only **(Fig. 2)** and one for pitching statistics only **(Fig. 3)**. In **Fig. 2**, we notice that the correlations with $r > 0.5$ are RBI, R, AvgRunPerGame, SLG, and OPS. In **Fig. 3**, the strongest correlations are ARAPG, RAllowed, WHIP, SV, and HAllowed. These were a part of our decision even though we used stepwise algorithms to see which combinations of variables were the best for final models.

Methods

With preliminary models, our intercept was very strange. Essentially, the intercept could only be interpreted if a team were to allow zero home runs, get zero hits, have a slugging statistic of zero, etc. None of these scenarios are realistic, so we instead centered each possible predictor variable at its median to allow our intercept to be meaningful.

Additionally, when considering all combinations of predictor variables, there are 2^{32} possible models, or almost 4.3 billion. This immediately forced us to use a stepwise algorithm to find the "best" combinations. The drawback of this method is that we don't get to examine the best combinations for models that have 1, 2, . . . , 32 total predictors; we instead just get the "best" model overall.

Results

We chose to split the models into three groups: Batting only, Pitching only, and both Batting and Pitching. This was done because batting is purely offensive, whereas pitching is purely defensive. Finally, we combined both to see overall, what predictors make a Twins win more likely. It is important to note that upon assessing each of our models with a residual plot, a quantile-quantile plot, and a histogram of residuals, no significant issues were found. The stepwise algorithm found the following models:

Batting Statistics ONLY:

$WinPer\hat{c}entage = 49.36 - 0.07\,(AB\,\hat{c}enter) + 0.09\,(H\,\hat{c}enter) + 0.10\,(HR\,\hat{c}enter)$

The initial model selected by the stepwise model for batting statistics contained no insignificant terms, thus we keep this model in its entirety.

Pitching Statistics ONLY:

$WinPer\hat{c}entage = 56.71 - 9.94(ARAP\hat{G}\ center) + 0.48(SV\ \hat{c}enter) + 0.90(HRAllo\hat{w}ed\ center) - 127.43(HR9\ \hat{c}enter)$

The initial model chosen by the stepwise algorithm had two predictors that were statistically insignificant, namely $HR9$ *center* and *HRAllowed center*. It turns out that upon dropping $HR9$ *center*, *HRAllowed center* becomes significant. This somewhat makes intuitive sense, as $HR9$ *center* is simply a per nine inning average of *HRAllowed center*. This leave us with:

$WinPer\hat{c}entage = 49.26 - 10.72(ARAP\hat{G}\ center) + 0.53(SV\ \hat{c}enter) + 0.10(HRAllo\hat{w}ed\ center).$

Pitching and Batting Statistics:

$$\hat{WinPercentage} = 49.35 - 10.63(ARAP\hat{G}\ center) + 0.22(\hat{SV\ center}) + 3.69(\hat{AvgRunPerGame\ center}) - 0.01(\hat{BB\ center}) - 0.05(\hat{HR\ center}) + 0.01(\hat{BBAllowed\ center}) + 158.18(\hat{SLG\ center}) + 2.85(\hat{HR9\ center})$$

The combined model initially selected by the algorithm had several insignificant terms. These insignificant terms were $HR9\ center$, $BBAllowed\ center$, and $BB\ center$. To ensure the overall effectiveness of the model, we dropped them one at a time in the order that they appeared in the stepwise algorithm output whilst checking to see if the remaining predictors became insignificant. Fortunately, upon dropping $HR9$, $BBallowed$, and BB from the model, all other predictors remain significant. This left us with:

$$\hat{WinPercentage} = 49.22 - 9.21(ARAP\hat{G}\ center) + 0.20(\hat{SV\ center}) + 3.60(\hat{AvgRunPerGame\ center}) - 0.04(\hat{HR\ center}) + 158.61(\hat{SLG\ center})$$

Interpretation

Since all of our predictors across all models are numeric only, interpretations will be the same across all predictors. For example, if one wanted to interpret the coefficient of $ARAP\hat{G}\ center$ from the combined model, one would say that ceteris paribus, on average, for every 1 additional run allowed per game above the median (4.07), that corresponds with the Twins' win percentage decreasing by 9.21 percentage points.

Discussion

Models and Implications

Looking at the batting only model, At Bats (AB), Hits (H), and Home Runs (HR) are the best predictors of a Twins win for their offense. Hits and home runs make sense as great predictors since more hits and home runs especially will result in a higher score.

A strange statistic which appeared in our batting model was AB, which does make sense at a second glance. If a team has more at bats, they are cycling through more batters and those runners are getting on base rather than out. However, the coefficient is negative, meaning more at bats would equal a lower chance of winning. Moving onto the pitching only model, Average runs allowed per game (ARAPG), Saves (SV), and Home Runs Allowed (HRAllowed) are the best predictors for a Twins win based solely on defense. Average runs allowed per game being strong makes sense since more allowed runs would mean lower win probability. Saves also check out as each save is directly responsible for a win. However, there was some trouble with Home Runs Allowed. The coefficient is positive, meaning more home runs allowed contributes to a higher win percentage, which doesn't make logical sense in the context of baseball. Lastly, the combined model tells us that the best predictors are Average runs allowed per game (ARAPG), Saves (SV), Average runs per game (ARPG), Home Runs (HR), and Slugging Percentage (SLG). Average runs allowed per game and saves both have positive coefficients just as they did in the other two models. Average runs per game and slugging percentage both have positive coefficients. This makes sense as an increase in either one means that more runners are getting on base for the Twins. Once again, there is an odd predictor, home runs. The coefficient for home runs is negative, meaning the more home runs the Twins hit, the lower their win percentage becomes, which again does not make sense in baseball. In hindsight, this is possibly due to the drawbacks of the stepwise selection algorithm.

Confounding Variables

Upon testing for multicollinearity by looking at variance inflation factors (VIF), we found that SLG had a VIF score of 15. This abnormally high, but when other predictors in the combined model are considered, they are also directly related to the Twins putting runners on base, specifically Home Runs and Average runs per game.

Was 2023 an outstanding year for the Twins?

According to all three of our models, 2023 was not an outlier year whatsoever. The $\hat{Win\ Percentage}$ fell well within the prediction intervals for all three models. However, this does not mean that 2023 was an ordinary year by any means.

The 2023 season saw the Twins set organization records in the following categories: Strikeouts, Fielding Percentage (FLD), Strikeouts Pitched (SOPitched), WHIP (Lowest in 50 years), and Errors. We found this by sorting the dataset by each predictor variable and observing if 2023 was on top of the resulting list. We did this for all years from 1901 to 2023; Therefore, with this in mind, we can conclude that the Twins defense and pitching has been the strongest it has ever been in organization history. It is plausible to say that the Twins' defense was a cause of their recent success. It is worth noting that sports, even ones with as many recorded statistics as baseball, are notoriously difficult to predict.

References

1. “At Last! Twins Snap 18-Game Postseason Losing Streak.” MLB.Com, https://www.mlb.com/news/twins-win-game-1-al-wild-card-series-2023. Accessed 13 Nov. 2023.

2. "DiamondDebates: MLB Strike | Baseball Hall of Fame." https://baseballhall.org/discover-more/stories/whole-new-ballgame/mlb-strike. Accessed 4 Dec. 2023.

3. “How Baseball Has Changed since 1908.” MLB.Com, https://www.mlb.com/news/baseball-has-changed-drastically-since-1908-c206845956. Accessed 13 Nov. 2023.

4. “Minnesota Twins Team Yearly Batting Stats.” Baseball-Reference.Com, https://www.baseball-reference.com/teams/MIN/batteam.shtml. Accessed 13 Nov. 2023.

5. “Minnesota Twins Team Yearly Pitching Stats.” Baseball-Reference.Com, https://www.baseball-reference.com/teams/MIN/pitchteam.shtml. Accessed 13 Nov. 2023.

6. “‘This Is My Dream’: How Royce Lewis Embodies the Upstart Twins.” ESPN.com, 10 Oct. 2023.

Appendix

Fig. 1: Variable Table

Variable Name	Explanation
ARAPG	Average runs allowed per game
SV	Saves - see Data and Variables section
HRAllowed	Home runs allowed
HR9	Home runs allowed on average per 9 innings
ARPG	Average runs per game
BB	Balls
BBAllowed	Balls allowed
SLG	Slugging - see Data and Variables section

Fig. 2: Batting Correlation Matrix

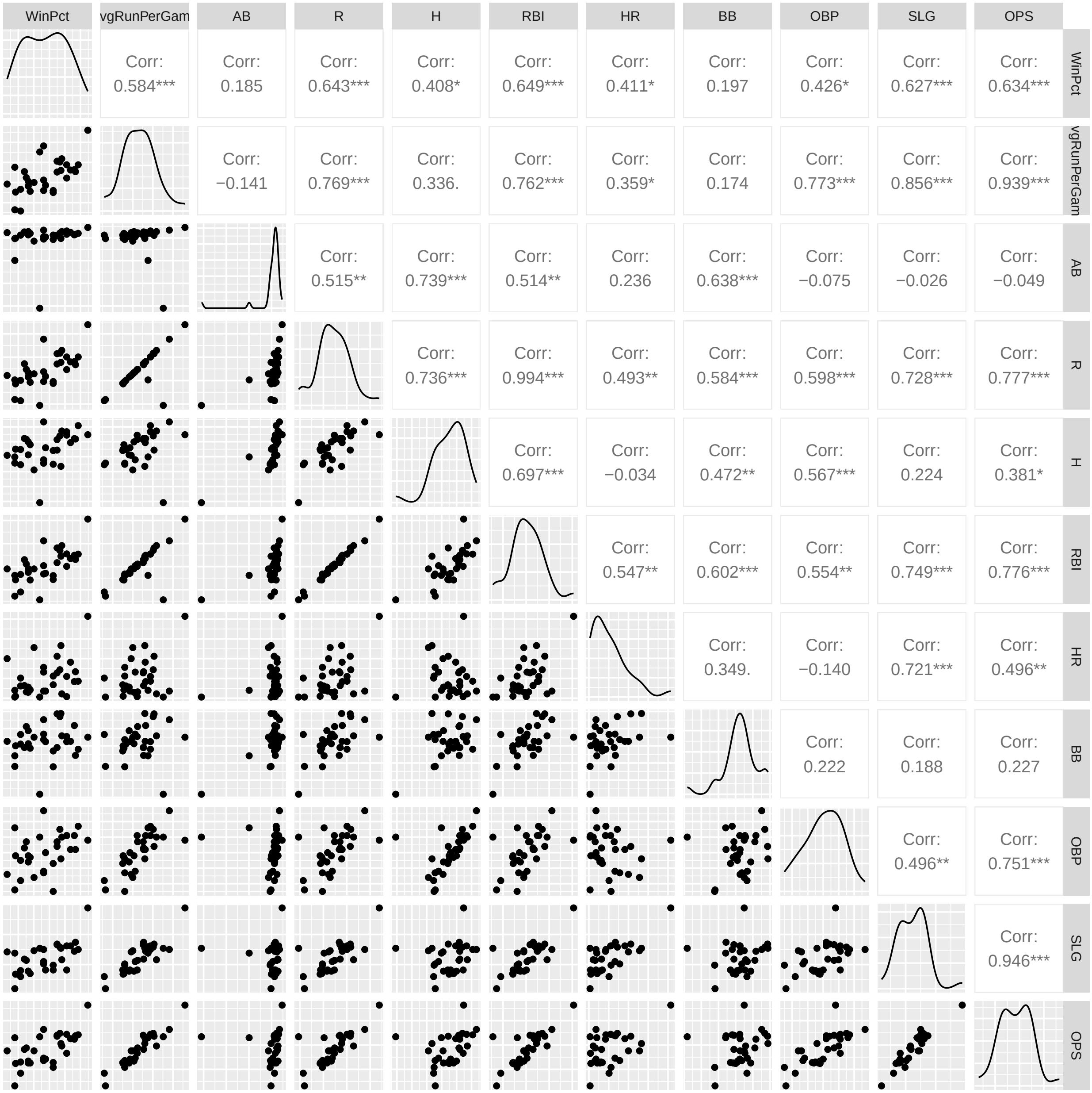
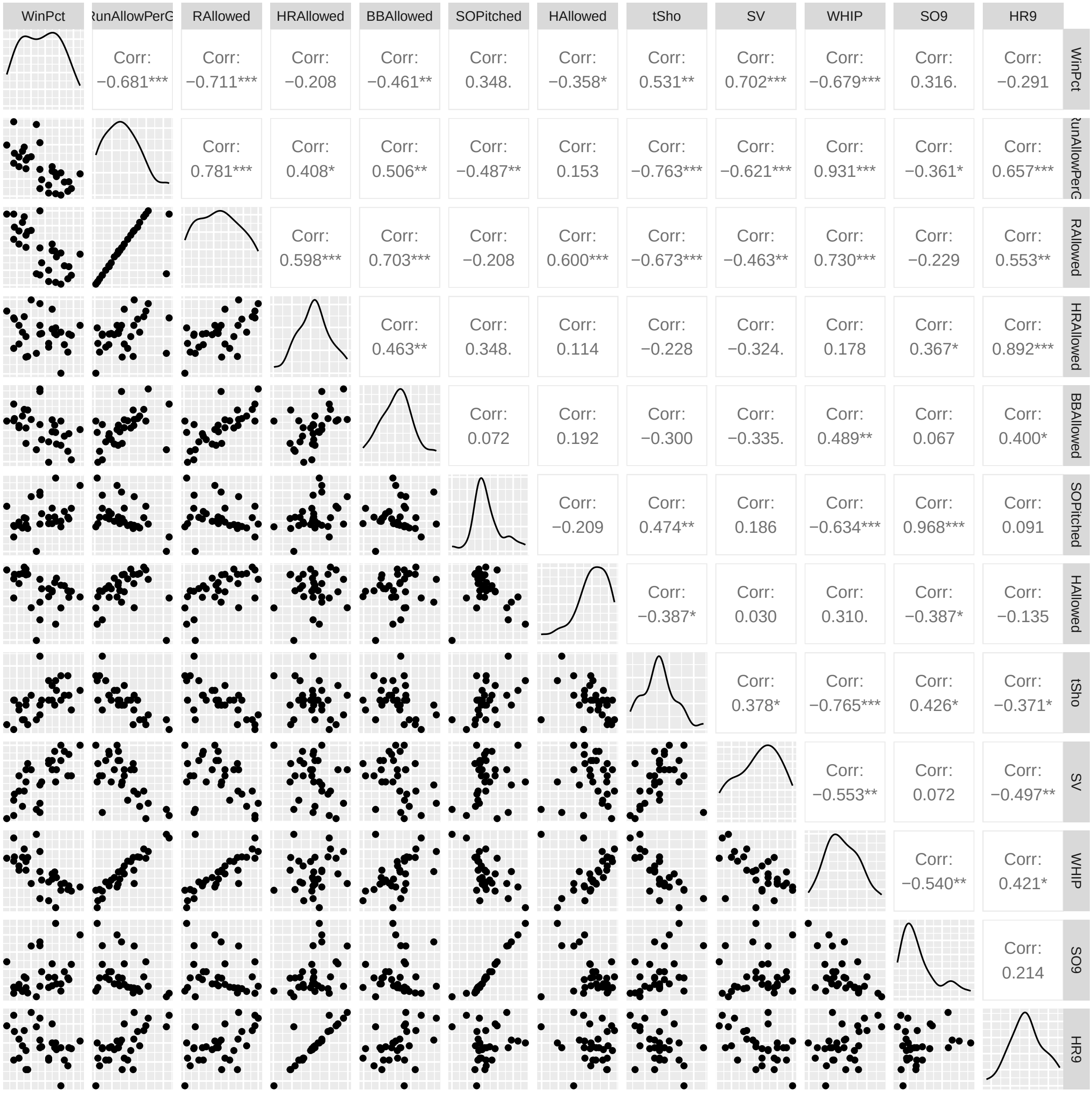


Fig 3: Pitching Correlation Matrix



CIR Visit Summary

We visited with Will Brandt in the CIR. He informed us that we should center our predictors as well as provide information for cryptic variable names.