Overall Context for Project

For over 100 years after the inception of professional baseball in America, Pitchers have been evaluated based on whether their team won the game and how many runs the team gave up, both of which are extremely dependent on the rest of the team. That finally changed in the late 1990s when Voros McCracken's research uncovered that a pitcher's batting average on balls in play had no consistency from year to year. This implied that a pitcher had little control over whether a ball in play was a hit or an out. The only pitching statistics that were correlated from year to year were home runs allowed, walks and strikeouts, which became known as the three true outcomes. That gave rise to Fielding Independent Pitching, or FIP, which is a pitching statistic designed to be on the same scale as Earned Run Average, or ERA, but only takes into account the three true outcomes. Over time, more advanced pitching statistics evolved that used the three true outcomes to try to evaluate a pitcher's performance. One of these is called SIERA, which accounts for more complex trends involving the three true outcomes and incorporates whether a batted ball was a grounder, popup or fly ball as well. One example of this is that it includes a walks squared term because pitchers that have high walk rates increasing their walk rate slightly means more runs will score than pitchers with low walk rates increasing their walk rate slightly because the pitchers with high walk rates are more likely to already have runners on base. Here is a link to a description of the different trends SIERA incorporates into its equation: https://blogs.fangraphs.com/new-siera-part-two-of-five-unlocking-underrated-pitching-skills/.

Problem Definition

Like the other metrics described above, my goal is to create a metric that is on the same scale as ERA but uses the three true outcomes. Similar to SIERA, I want to incorporate whether a batted ball was a grounder, popup or fly ball as well. Pitchers are thought to at least have some level of influence on this as opposed to whether a batted ball is a hit or an out. In addition, I want to incorporate exit velocity into my model. This is another metric for which it's questionable how much control a pitcher has in its outcome, but it would make sense that a pitcher has more control over it than whether a batted ball is a hit because it's directly representative of the quality of contact whereas a softly hit ball may drop for a hit while a hard line drive hit right at a fielder could be an out.

Project Motivation

Major League Baseball set a record revenue in 2022 of nearly $11 billion. Any incremental increase in ability to evaluate players better could be extremely valuable. For example, the Oakland Athletics of the early 2000s had a budget that was a fraction the size of their competitors but they still won over 100 games in multiple seasons due to their superior evaluations of players using data analytics.

Proposed Methodology

My plan is to use linear regression to create an equation on the same scale as ERA that incorporates the outcomes of which pitcher is believed to have the most control. These features include the three true outcomes, whether a batted ball was a grounder, popup or flyout, combinations of two of these variables multiplied together as used in SIERA (i.e. walks squared), and average exit velocity, with ERA as the target variable. I will use forward stepwise regression to select the features. This process starts with an empty model and adds features to the model until the adjusted R squared peaks. This process allows the machine to select the best features rather than relying on our assumptions of which of them are valuable.

Proposed Data Source

MLB's Baseball Savant tool (https://baseballsavant.mlb.com/leaderboard/statcast?type=pitcher) has a variety of pitching statistics, including exit velocity and allows users to easily download their data as a CSV file. Additional data sources may be added as needed.

Potential Real World Applications

As mentioned, any increase in ability to evaluate players is extremely valuable to teams that spend hundreds of millions of dollars a year on player payroll. In addition, fans use stats like these to evaluate pitchers. While WAR, the most popular baseball cumulative value statistic, is very in depth for hitters, for pitchers it simply uses a cumulative form of either RA/9, the number of runs a pitcher allows per nine innings, or FIP. A better metric of pitcher performance on a per nine innings basis could make WAR a better cumulative value statistic for pitchers.