***Baseball Key Predictors for 2024 Utilizing Linear Regression***

**By: Jordan S Richman**

**Introduction:**

For every offseason since the end of the 2022 season I’ve been developing an algorithm to accurately predict the key contributors for each team and try to predict the Most Valuable Player for both the National and American League (excluding pitchers as it’s unlikely pitchers exceed beyond the Cy Young). For my analysis for the 2024 season, I utilized 2021, 2022, and 2024 data where I particularly focused on analyzing a players expected statistics instead of their based statistics (Batting Average, OPS, etc) and counting statistics (Hits, Home Runs). I wanted to capture a wide range of players thus I limited the minimum needed plate appearance for my analysis at 100 PA’s. I decided to limit the data to only to the 2021 season as the 2020 season was shorten and a vast number of players did not play and players tend to pass through the Major League Baseball system quickly as it’s very difficult to maintain a position. Throughout my analysis you will see I utilized Machine Learning algorithms on various advanced expected statistics and then assign player points based on where they stand in the top ten of their team since part of their MVP contributions is how much value they hold to their own team. Linear Regression was performed on these statistics to provide us a predicted statistic on how well a player should perform in the coming year.

**Definitions:**

**Most Valuable Player:** An annual Major League Baseball award given to one outstanding player in the American League and one in the National League. The award has been presented by the Baseball Writers' Association of America since 1931.

**Expected Statistics:** “xStats are calculated using Statcast data in an attempt to make more objective observations of the game. Only the vertical and horizontal launch angles, exit velocities, batted ball distances, game time temperature, and ball park are taken into account. All other factors are ignored. The angle and exit velocity information is fed through an algorithm that lumps together similarly hit balls and finds their average success rates. Game time temperature and ball park information are used to adjust the exit velocities.”- Andrew Perpetua, creator of xStats.

**Plate Appearance:** In baseball, a player is credited with a plate appearance each time he completes a turn batting. Under Rule 5.04 of the Official Baseball Rules, a player completes a turn batting when he is put out or becomes a runner.

**Hit**: A hit occurs when a batter strikes the baseball into fair territory and reaches base without doing so via an error or a fielder's choice. There are four types of hits in baseball: singles, doubles, triples and home runs. All four are counted equally when deciphering batting average. If a player is thrown out attempting to take an extra base (e.g., turning a single into a double), that still counts as a hit.

**Homerun:** A home run occurs when a batter hits a fair ball and scores on the play without being put out or without the benefit of an error.

**Expected Weighted On-base Average (xwOBA)** - xwOBA is formulated using exit velocity, launch angle and, on certain types of batted balls, Sprint Speed. In the same way that each batted ball is assigned an xBA, every batted ball is given a single, double, triple and home run probability based on the results of comparable batted balls since Statcast was implemented Major League wide in 2015. xwOBA also factors in real-world walk and strikeout numbers, and is reported on the wOBA scale. By comparing expected numbers to real-world outcomes over a period of time, it can be possible to identify which hitters (or pitchers) are over- or under-performing their demonstrated skill.

**Expected Batting Average (xBA):** xBA measures the likelihood that a batted ball will become a hit. Each batted ball is assigned an xBA based on how often comparable balls -- in terms of exit velocity, launch angle and, on certain types of batted balls, Sprint Speed -- have become hits since Statcast was implemented Major League wide in 2015. By comparing expected numbers to real-world outcomes over a period of time, it can be possible to identify which hitters (or pitchers) are over- or under-performing their demonstrated skill.

**Barrels:** A batted ball with the perfect combination of exit velocity and launch angle, or the most high-value batted balls. (A barrel has a minimum Expected Batting Average of .500 and Expected Slugging Percentage of 1.500.)

**Launch Angle (LA):** How high/low, in degrees, a ball was hit by a batter.

**Exit Velocity (EV)(aka Hit Speed):** How fast, in miles per hour, a ball was hit by a batter.

**Slugging Percentage (SLG):** Slugging percentage represents the total number of bases a player records per at-bat. Unlike on-base percentage, slugging percentage deals only with hits and does not include walks and hit-by-pitches in its equation.

**On-base Percentage (OBP):** OBP refers to how frequently a batter reaches base per plate appearance. Times on base include hits, walks and hit-by-pitches, but do not include errors, times reached on a fielder's choice or a dropped third strike. (Separately, sacrifice bunts are removed from the equation entirely, because it is rarely a hitter's decision to sacrifice himself, but rather a manager's choice as part of an in-game strategy.)

**On-base Plus Slugging (OPS):** OPS adds on-base percentage and slugging percentage to get one number that unites the two. It's meant to combine how well a hitter can reach base, with how well he can hit for average and for power.

**Expected Slugging Percentage (xSLG):** formulated using exit velocity, launch angle and, on certain types of batted balls, Sprint Speed.

**Isolated Power (ISO):** Measures the raw power of a hitter by taking only extra-base hits -- and the type of extra-base hit -- into account.

For example, a player who goes 1-for-5 with a double has an ISO of .200. A player who goes 2-for-5 with a single and a double has a higher batting average than the first player, but the same ISO of .200.

**Fielding Run Value (FRV):** is Statcast's overall metric for capturing a player’s measurable defensive performance onto a run-based scale, which can then be read as a player being worth X runs above or Y runs below average. Since different types of defensive performance are expressed in different scales -- throws, outs, blocks, etc. -- this conversion is necessary to place all performance on the same scale, which then allows defenders of all positions to be compared to one another.

**Body:**

**xwOBA Regression Model-**

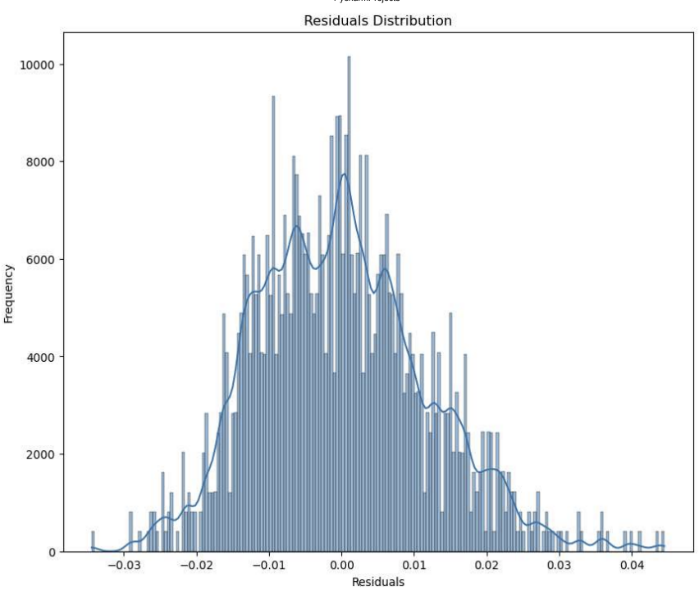
A diagram of a graph

Description automatically generated with medium confidence

The first statistic I wanted to do analysis on was xwOBA or Estimated Weighted On Base Average. xwOBA is an advanced metric that predicts a player’s On Base Average utilizing various factors such as ballpark, launch angle, etc. to give us a more detailed and accurate result. I constructed a Linear Regression model utilizing xwOBA data from 2021 – 2023 to make an estimated predicted on a player’s xwOBA for 2024. The model indicates the League Average xwOBA at 0.320 and the 80th percentile xwOBA at 0.340. The R2 for this model is fairly high at 0.90 which indicates that a large portion of the variability in the predicted 2024 xwOBA is explained by the variability in the xwOBA from 2021 to 2023. The Mean Square Error for this model is fairly close to 0 at 0.000139 indicates that the model's predictions are close to the actual values, and the model is performing well in terms of minimizing the squared differences.

Additionally, I wanted to display the top ten player’s and the bottom five players in terms of xwOBA. Some of the more notable players are Yordan Alvarez (appearing twice since he placed high in xwOBA in multiple years), Bryce Harper, and Byron Buxton. I created a new variable labeled “predicted\_est\_woba(advanced)” that shows a players estimated xwOBA for 2024. It’s important to note that this model estimates using only 100 Plate Appearances so even a player plays well for a short period of time they could place in the higher percentiles. An example of this is Byron Buxton. Buxton is a great hitter, but in most seasons, he gets injured halfway through the season and misses a large portion of the season. In the past three seasons he has averaged 293 At Bats which is fairly low for an All-Star caliber player.

A diagram of a graph showing a blue dotted line

Description automatically generated with medium confidence

Given our R-squared in the previous model, we can easily calculate our R-value by square rooting our R-squared to give us our R-value of 0.9487. Since were looking for an R-value as close as possible to 1 on a scale from -1 to 1, we can conclude that our R-value of 0.9487 indicates a strong positive linear relationship.

The next two following plots are utilized to test our model’s predictive performance. The first shows a scatter plot between Residuals and Predicted 2024 xwOBA where there’s no underlaying patterns that has a random scatter around 0. In which is a positive sign for the model’s predictive performance. The other plot is a histogram of the Residuals which resembles a normal distribution (normal bell curve). Most of the data falls near the 0 Residual mark and the rest of the data flows normally across the bell curve. This helps ensure our model is not systematically underestimating or overestimating the target xwOBA. Since both plots are favorable, we can conclude that our Regression Model is suitable for our analysis.

A graph of blue color

Description automatically generated with medium confidenceLastly, since xwOBA is such a well-rounded statistic to estimate a player’s performance. I created a a bar chart that shows top 30 Players with the Highest Predicted xwOBA for 2024. These players are likely to produce (barring any injuries) at a higher level than the majority of the Major League. These players will get on base and hit the ball at the advanced level hence why most of these players are not any surprise players since they have established themselves.

**xBA Regression Model-**

A diagram of a graph

Description automatically generated with medium confidence The next statistic I analyzed was xBA or Estimated Batting Average. Batting average has lost its value in the recent years, however xBA provides a more rounded statistic that better shows a player’s hitting ability. xBA is an advanced metric that predicts a player’s Batting Average utilizing various factors such as ballpark, launch angle, etc. to give us a more detailed and accurate result. I constructed a Linear Regression model utilizing xBA data from 2021 – 2023 to make an estimated predicted on a player’s xBA for 2024. The model indicates the League Average xBA at 0.248 and the 80th percentile xBA at 0.264. The R-squared for this model is high at 0.86 which indicates that a large portion of the variability in the predicted 2024 xBA is explained by the variability in the xBA from 2021 to 2023. The Mean Square Error for this model is close to 0 at 0.00010 indicates that the model's predictions are close to the actual values, and the model is performing well in terms of minimizing the squared differences.

Additionally, I wanted to display the top ten player’s and the bottom five players in terms of xBA. Some of the more notable players are Aaron Judge (appearing twice since he placed high in xBA in multiple years), Ronald Acuna Jr, and Corey Seager. I created a new variable labeled “predicted\_est\_ba” that shows a players estimated xBA for 2024. It’s important to note that this model estimates using only 100 Plate Appearances so even a player plays well for a short period of time they could place in the higher percentiles as we noted in a previous analysis.

A diagram of a graph showing a number of blue dots

Description automatically generated with medium confidence A graph showing a distribution of residuals

Description automatically generated

Given our R-squared in the previous model, we can easily calculate our R-value by square rooting our R-squared to give us our R-value of 0.9274. Since were looking for an R-value as close as possible to 1 on a scale from -1 to 1, we can conclude that our R-value of 0.9274 indicates a strong positive linear relationship.

The next two following plots are utilized to test our model’s predictive performance. The first shows a scatter plot between Residuals and Predicted 2024 xBA where there’s no underlaying patterns that has a random scatter around 0. In which is a positive sign for the model’s predictive performance. The other plot is a histogram of the Residuals which resembles a normal distribution (normal bell curve). Most of the data falls near the 0 Residual mark and the rest of the data flows normally across the bell curve. This helps ensure our model is not systematically underestimating or overestimating the target xBA. Since both plots are favorable, we can conclude that our Regression Model is suitable for our analysis.

**xSLG Regression Model-**

A diagram of a graph

Description automatically generated with medium confidence